Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

October 2011

Meta-heuristic Strategies in Scientific Judgment

Spencer P. Hey
University of Western Ontario

Supervisor
Charles Weijer
The University of Western Ontario

Graduate Program in Philosophy
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy

Recommended Citation

Hey, Spencer P., "Meta-heuristic Strategies in Scientific Judgment" (2011). *Electronic Thesis and Dissertation Repository*. 292.
https://ir.lib.uwo.ca/etd/292

META-HEURISTIC STRATEGIES IN SCIENTIFIC JUDGMENT
(Thesis format: Monograph)

by

Spencer Hey

Graduate Program in Philosophy

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

THE UNIVERSITY OF WESTERN ONTARIO
School of Graduate and Postdoctoral Studies

CERTIFICATE OF EXAMINATION

Supervisor:

. . . .
Dr. C. Weijer

Supervisory Committee:

. . . .
Dr. R. Batterman

. . . .
Dr. G. Barker

Examiners:

. . . .
Dr. C. Smeenk

. . . .
Dr. G. Barker

. . . .
Dr. W. Wimsatt

. . . .
Dr. M. Speechley

The thesis by

Spencer Phillips Hey

entitled:

Meta-heuristic Strategies in Scientific Judgment

is accepted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

. . . .
Date

. . . .
Chair of the Thesis Examination Board

Abstract

In the first half of this dissertation, I develop a heuristic methodology for analyzing scientific solutions to the problem of underdetermination. Heuristics are rough-and-ready procedures used by scientists to construct models, design experiments, interpret evidence, etc. But as powerful as they are, heuristics are also error-prone. Therefore, I argue that they key to prudently using a heuristic is the articulation of meta-heuristics—guidelines to the kinds of problems for which a heuristic is well- or ill-suited.

Given that heuristics will introduce certain errors into our scientific investigations, I emphasize the importance of a particular category of meta-heuristics involving the search for robust evidence. Robustness is understood to be the epistemic virtue bestowed by agreement amongst multiple modes of determination. The more modes we have at our disposal, and the more these confirm the same result, the more confident can we be that a result is not a mere artifact of some heuristic simplification. Through an analysis of case-studies in the philosophy of biology and clinical trials, I develop a principled method for modeling and evaluating heuristics and robustness claims in a qualitative problem space.

The second half of the dissertation deploys the heuristic methodology to address ethical and epistemological issues in the science of clinical trials. To that end, I develop a network model for the problem space of clinical research, capable of representing the various kinds of experiments, epistemic relationships, and ethical justifications intrinsic to the domain. I then apply this model to ongoing research with the antibacterial agent, moxifloxacin, for the treatment of tuberculosis, tracking its development from initially successful and promising in vitro and animal studies to its disappointing and discordant performance across five human efficacy trials. Given this failure to find a robust result with moxifloxacin across animal and human studies, what should researchers now do? While my final analysis of this case does not definitively answer that question, I demonstrate how my methodology, unlike a statistical meta-analysis, helps to clarify the directions for further research.

Keywords: Heuristics, robustness, epistemology, methodology, clinical trials, Duhem

Contents

List of Figures

List of Tables

Chapter 1

Duhem's Solution

Pierre Duhem observed that even the best designed experiments never test hypotheses in isolation. It is the experiment as a whole—a combination of theories, hypotheses, and auxiliary assumptions—that confronts the evidence. A recalcitrant experimental result is therefore never enough, on its own, to provide a conclusive refutation of any single hypothesis; evidence is never sufficient to prescribe how a theory ought to be constructed or revised. So how does a scientist go about modifying her theories? And more importantly, how *should* she go about it?

Philosophers of science have made much of this "Duhem problem". However, most have overlooked the fact that Duhem also proposed a solution.[1] Scientific judgments are underdetermined, but it does not follow that they must be irrational or arbitrary. For Duhem, underdetermination simply meant that judgments cannot be made on the basis of logic. Instead, they must rely on what he calls "good sense".

The positivist's distinction between the context of discovery and the context of justification, combined with the fact that Duhem does not say very much about "good sense" in his *Aim and Structure of Physical Theory*, provides a plausible explanation for philosophers overlooking his solution. "Good sense", whatever it may be, is only relevant to the context of discovery, which is assumed to be irrational and therefore irrelevant to philosophical reflection. Although this dubious distinction in contexts is no longer as dominant in the philosophy of science, a rigorous analysis of the "intuitive" judgments contributing to scientific practice; judgments fulfilling the promises of Duhemian good sense, has yet to appear. The aim of this thesis is to provide exactly this rigorous analysis: What is Duhemian good sense? How does it function in

[1] It is worth noting that a persistent state of confusion surrounds philosophical discussion of "Duhem's problem" or "the Duhem thesis". Much of the attention paid to this subject in the 20th century confuses Duhem's ideas with those of W. V. O. Quine, substituting the former's rather modest form of underdetermination for versions of the latter's more radical holism. Although there are now a number of excellent and clarifying responses to this confusion (cf. Laudan 1965, Ariew 1984, Needham 2000, Darling 2002), it bears repeating that the radical "Duhem-Quine Thesis", that hypotheses can be maintained in the face of any evidence whatsoever, is not a position Duhem held.

scientific practice? How can we develop the concept further to aid in scientific progress?

But I should be clear: This is not a historical project. Duhem's ideas about underdetermination and good sense will largely serve as a philosophical touchstone, an epistemological problem that will be familiar to many philosophers of science. The next section will be the most historical, as I will show that Duhem's concept of good sense is not mere hand-waving, but is fundamental to his picture of scientific practice. But the majority of my work is going well beyond Duhem's basic idea. Specifically, I will argue that we can elucidate his conception of good sense with contemporary work on heuristics. Heuristics are rough (and usually simple) problem-solving procedures, well-adapted to certain kinds of problems. The complexity of virtually all scientific phenomena demands that scientists use heuristics in their everyday practice. Therefore, I propose that we can think of good sense as judgments about the appropriate use of particular heuristics for particular problems. But I am not arguing that my account is exactly what Duhem had in mind, or even that he would necessarily agree with my extension of his ideas. I am instead arguing that his project of developing good sense is a fruitful starting point for a renewed investigation into the practical tools of scientific judgment.

I begin, in §1.1, by examining Duhem's writing on good sense. Although he clearly recognizes its central importance for scientific judgments of many kinds, Duhem does not take good sense beyond the notion of an "intuitive mind". Fortunately, as I have suggested above, we can now go further than this and say something more precise about the character of scientific "intuition".

In §1.2, I will review one proposal for "solving" Duhem's problem and making the techniques of scientific judgments more precise. Dorling (1979) and Howson and Urbach (1993) both put forward Bayesian confirmation theory as a solution to Duhem's problem. Although I argue that their (essentially identical) accounts are inadequate, their failure is nevertheless instructive for drawing our attention to the specific judgments relevant to Duhem's problem.

Following that, in §1.3, I propose my elucidation of scientific judgments as heuristics—rough-and-ready rules of simplification and interpretation. I argue that the prescriptions of good sense can be understood as *meta-heuristics*, guidelines and strategies for the fruitful application of scientific heuristics, governing a problem space.[2] I then illustrate my account with examples of mathematical explanations in physics, decision-making in clinical medicine, and dynamic modeling in ecological management. Each of these cases provides a contrasting set of heuristics and lessons about the "regions" of problem space appropriate for their application.

I conclude, in §1.4, with a summary of this chapter's main points and a brief sketch of the

[2] As I use the term, a "problem space" is a conceptual way to talk about the set of all possible solutions to a problem. Throughout the thesis I will discuss different ways of representing this "space" in order to understand decision-making practices in science.

investigations to come in the later chapters.

1.1 What is Good Sense?

Duhem's *Aim and Structure of Physical Theory* (1906) is the most familiar of his works for philosophers. It is there that he articulates his eponymous problem:

> When certain consequences of a theory are struck by experimental contradiction, we learn that this theory should be modified but we are not told by the experiment what must be changed. It leaves to the physicist the task of finding out the weak spot that impairs the whole system. No absolute principle directs this inquiry, which different physicists may conduct in very different ways without having the right to accuse one another of illogicality (Duhem 1906, p.216).

His point here is quite simple: The "absolute principles" of logic cannot guide theoretical revisions.[3] Therefore, different scientists may employ different strategies:

> One [physicist] may be obliged to safeguard certain fundamental hypotheses while he tries to reestablish harmony between the consequences of the theory and the facts by complicating the schematism in which these hypotheses are applied, by invoking various causes of error, and by multiplying corrections. The next physicist, disdainful of these complicated artificial procedures, may decide to change one of the essential assumptions supporting the entire system. The first physicist does not have the right to condemn in advance the boldness of the second one, nor does the latter have the right to treat the timidity of the first physicist as absurd. The methods they follow are justifiable only by experiment, and if they both succeed in satisfying the requirements of experiment each is logically permitted to declare himself content with the work that he has accomplished (Ibid., pp.216-217).

If this were all Duhem had said on the matter of underdetermination, there would be little reason to object to equating his ideas with radical holism. However, it is important to see that he qualifies his comments with the phrase "logically permitted". Logical underdetermination is no stronger than, and entirely consistent with, the first passage above. If one scientist wishes to consider a hypothesis refuted while the other wishes to revise the underlying theory, then logic will accommodate them both. For the purpose here, the critical point is what immediately follows:

[3] Although Duhem refers here only to physicists, we will soon see that his views on underdetermination and good sense are not limited to physics.

> This does not mean that we cannot very properly prefer the work of one of the two to that of the other. Pure logic is not the only rule for our judgments; certain opinions which do not fall under the hammer of the principle of contradiction are in any case perfectly unreasonable. These motives which do not proceed from logic and yet direct our choices, these "reasons which reason does not know" and which speak to the ample "mind of finesse" but not to the "geometric mind", constitute what is appropriately called good sense (Ibid., p.217).

It is remarkable that with all the attention given to Duhem's problem, so little philosophical attention has been paid to this passage.[4] The use of good sense—consisting of "reasons which reason does not know"—is Duhem's solution to the problem of underdetermination. Far from suggesting that any revisions must be irrational or arbitrary, he states that we have resources to "very properly prefer" some corrective methods over others. Yet again, these resources are outside the realm of logic, and as a result, scientists may very well disagree about their prescriptions:

> [The] reasons of good sense do not impose themselves with the same implacable rigor that the prescriptions of logic do. There is something vague and uncertain about them; they do not reveal themselves at the same time with the same degree of clarity to all minds. Hence, the possibility of lengthy quarrels between the adherents of an old system and the partisans of a new doctrine, each camp claiming to have good sense on its side, each party finding the reasons of the adversary inadequate. The history of physics would furnish us with innumerable illustrations of these quarrels at all times and in all domains. . . . In any event this state of indecision does not last forever. The day arrives when good sense comes out so clearly in favor of one of the two sides that the other side gives up the struggle even though pure logic would not forbid its continuation (Ibid., pp.217-218).

This passage adds another layer to the concept of good sense, and suggests that it is more than a tool of the individual scientist's judgment. Scientists may disagree in the short run about what are the prescriptions of good sense. Yet, Duhem thinks that eventually this disagreement will give way, succumbing to the weight of mounting experimental evidence.

Putting these few pieces together, we can now see that Duhemian good sense is a judgment about the particular way a scientist goes about revising her theories. In other words, it is a second-order, methodological judgment of the efficacy of a first-order judgment. Since neither of these judgments is determined by logic, scientists may differ about both the proper method and the prescriptions of good sense. But over time, Duhem thinks, it will become clear that

[4] The English translation of *Aim and Structure* (1991) from Princeton University Press does not even index the phrase "good sense".

some first-order judgments—e.g., to reject an hypothesis or revise the underlying theory—are better than others. So it is that good sense "comes out in favor" of one approach.

Of course, the interesting question is *how* good sense can accomplish this task. What is the nature of these judgments such that they can be so arranged and refined? Merely labeling something "good sense" and gesturing toward its role in science is not of much help in itself. Sensitive to this worry, Duhem suggests that the scientist (or the philosopher of science) ought to consciously develop the notion:

> Since logic does not determine with strict precision the time when an inadequate hypothesis should give way to a more fruitful assumption, and since recognizing the moment belongs to good sense, physicists may hasten this judgment and increase the rapidity of scientific progress by trying consciously to make good sense within themselves more lucid and more vigilant (Ibid., p.218).

The work of this entire thesis is in the spirit of Duhem's suggestion here—the importance of developing the notion of good sense as an aid to scientific progress. Unfortunately, Duhem has nothing more to say about good sense in *Aim and Structure*. For more, we must turn to his 1915 work, *German Science*.

In *German Science*, it becomes clear that good sense plays a deeper role in Duhem's picture of scientific methodology. Grounding more than the interpretations of experimental evidence and its impact on theory, good sense also applies to the intuitions underlying the acceptability of axioms or "first principles". Introducing this extension, Duhem quotes from Blaise Pascal:

> We know the truth not only by reason, but also by the heart, and it is through the latter sort of knowledge that we know first principles. . . . And it is upon such knowledge of the heart and of instinct that reason must rest, and base all its discourse. The heart feels that there are three dimensions in space, and that numbers are infinite. Reason then demonstrates that there are no two squared numbers one of which is twice the other. Principles are intuited, propositions are inferred; all lead to certitude, though by different routes (Pascal 1958).

Duhem then comments:

> Good sense, which Pascal calls the "heart" here, for the intuitive perception of the obviousness of the axioms, and the deductive method to arrive by the rigorous but slow progress of *discourse* at the demonstration of the theorems: there we have the two means that human intelligence uses when it wishes to construct a science of reasoning (Duhem 1915 p.8, original emphasis).

The distinction between logic and good sense (i.e., deduction and intuition) described in *Aim and Structure* is still evident here. What is new is an extension of good sense from its role in the interpretation of experiments to other intuitive judgments, specifically, the "obviousness of axioms".

Duhem is also explicit that good sense applies beyond the judgments of the physicist. It applies to the mathematician (in their choice of axioms), as well as, the biologist:

> From his preconceived idea an experimenter such as Louis Pasteur infers this consequence: if one injects one particular substance into rabbits, they will die. The control animals which have not had the injection will remain in good health. But this observer is aware that a rabbit might die, at times, from other causes than the injection whose effects are being studied. He knows equally that certain particularly resistant animals can sustain doses of an injection which would kill most of their fellows, or, further, than an inept administration can impair the injection and render it inoffensive. If, then, he sees one inoculated rabbit live or one control animal die, he need not conclude directly and overtly to the falsity of his preconceived idea. He could be faced with some accident of the experiment which need not require the abandonment of his idea. What will determine whether these failures are or are not of such a nature that the supposition in questions must be renounced? Good sense (Duhem 1915, pp.23-24).

And yet, should good sense prescribe that the preconceived idea ought to be renounced, its task is still not done:

> One must substitute for [the rejected idea] a new supposition which has the possibility of standing up better to experimental testing. Here it is necessary to pay attention to what each of the observations which have condemned the initial idea suggests, to interpret each of the failures which destroyed that idea, and to synthesize all these lessons for the purpose of fabricating a new thought which will pass once again under the scrutiny of the actual results. What a delicate task, concerning which no precise rule can guide the mind! It is essentially a matter of insight and ingenuity! (Ibid., p.24)

"Insight and ingenuity"—here we reach the end of what Duhem can teach us about good sense, and therefore, I will now begin shift my focus from the historical question, "what is Duhem's conception of good sense?" to the contemporary question, "how ought we to understand and develop good sense for use in current scientific practice?" While I share his view that "no precise rule can guide the mind" to make scientific judgments about first principles or the interpretation of experimental results, we need not limit ourselves to "insight and ingenuity".

Yet, before we leave Duhem, it is worth summarizing the scientific judgments to which he thinks good sense applies. This is the skeleton framework onto which any epistemic development rightly called "good sense" must be built. For Duhem, good sense is involved in four scientific activities:

1. Judging the "obviousness" of axioms and first principles, or construed more liberally, judging the acceptability of conceptual assumptions.

2. Interpreting experimental evidence in relation to theory, hypothesis testing, and auxiliary assumptions.

3. Developing new theories and hypotheses for evaluation.

4. Over time, resolving differences in scientific judgment.

Philosophical discussion of Duhem's *problem* has, by and large, been limited to the second of these. One advantage in focusing on Duhem's *solution* lies in placing these four activities side-by-side. No one questions that all of these things go on in science. Scientists make such judgments all of the time and the scientific community strives (most of the time) to resolve its internal disputes. Duhemian good sense brings them all together.

The question I want to consider for the remainder of this chapter is *how* good sense might fruitfully fulfill all of these obligations. How does a scientist decide which axioms or first principles to use? Or which areas of her theory need revision? Why does she judge some particular revisions to be better than others? Understanding the "intuitive" methods appropriate to resolving each of these difficulties; knowing how such methods work and when they will break down, is the key to elucidating the second-order judgments constituting Duhemian good sense.

As a philosophical exercise, this is both descriptive and prescriptive. The descriptive aspect of understanding good sense is identifying the judgment strategies themselves. When a scientist chooses to reject an hypothesis in light of evidence, why has she done so? What other factors did she consider? Or when a scientist constructs a mathematical model, why has he chosen to simplify or ignore some particular features of the system and not others? The answers to these questions will be descriptions of scientific practice. The prescriptive aspect of good sense is the meta-judgment: judging whether or not those judgments were the right ones to make.

In the next section, I will discuss the possibility of elucidating the descriptive and prescriptive elements of Duhem's solution using Bayesian confirmation theory. When Duhem writes of the limitations of logic to guide scientific judgments, he has in mind deductive logic. But perhaps inductive logic could be used to solve Duhem's problem and explain why a scientist revises her theory the way that she does.

1.2 The Duhem Problem "Solved" by Bayesian Means

In developing their accounts of scientific judgment, Dorling (1979) and Howson and Urbach (1993) rationally reconstruct historical examples to show how some particular judgment conforms to the prescriptions of the Bayesian epistemic calculus. To do this, they partition an actual scientist's supposed beliefs into a hypothesis, H, and a theory, T, only one of which can be accepted after a predictive failure. $T\&H$ imply some piece of evidence, E. If we observe $\neg E$, then from modus tollens, we must decide what amongst $T\&H$ we want to revise.

Dorling considers John Adams' mid-nineteenth century computation for the secular acceleration of the moon (E) in the context of established Newtonian theory (T) and the hypothesis (H) that the effects of tidal friction are negligible. The observational result (E') disagreed with his computation, so what should Adams have done? Revise T or H? (Dorling 1979, pp.178-179)

Howson and Urbach consider William Prout's hypothesis that the atomic weights of the elements are all whole-number multiples of the atomic weight of hydrogen (T). Observed measurements of chlorine's atomic weight (E') conflicted with Prout's prediction (in light of his hypothesis) (E), calling into question the accuracy of the measuring techniques of the time (H). So should Prout have revised T or H? (Howson and Urbach 1993, pp.97-98)

After plugging in quantitative values for a scientist's credence in T, H, and the various likelihoods needed to run the calculus, Dorling and Howson and Urbach are able to show that it was rational, in the Bayesian sense, for Adams and Prout to reject their respective H's rather than their respective T's. "Thus", conclude Howson and Urbach, "Bayes' Theorem provides a model to account for the kind of scientific reasoning that gave rise to the Duhem problem" (Ibid., p.101).

The review of Duhem's problem and solution in the previous section should make us doubt this hasty conclusion. What Dorling and Howson and Urbach show is that given a particular set of prior beliefs about a hypothesis and a theory, a recalcitrant result should differentially reduce our credence, leading us to posteriorly find the hypothesis very unlikely and the theory only slightly less probable than before. Despite the fact that they base their quantitative prior assignments on the respective scientists' (or scientific community's) historical comments, their demonstration is not actually a solution to Duhem's problem.

Recall that Duhem motivated his problem with the fact that two scientists may very well disagree about how to proceed, one "multiplying corrections" in the face of evidence, the other "changing one of the essential assumptions supporting the entire system". Nothing in these Bayesian accounts explains why we might still "very properly prefer the work of one of the two to that of the other". Should a different scientist choose an alternative hypothesis to revise, the Bayesian calculus would look the exactly the same.

For example, in Howson and Urbach's case, they point out that Prout doubted the accuracy of the measuring techniques for atomic weights at the time (H). In testing the atomic weight of chlorine, they argue that the recalcitrant result (E') of 35.83, rather than the 36 (E) predicted by Prout's "theory"—all atomic weights are whole number multiples of hydrogen's atomic weight (T)— explains and justifies this doubt. But their account does not provide anything about why Prout doubted H in the first place. Duhem's original point was exactly that such crucial experiments are impossible because the relevant hypotheses are underdetermined. By what reasoning did Prout come to suspect that the measuring techniques were unreliable? Dorling's and Howson and Urbach's accounts may serve to justify a decision about revising *some* hypothesis, once it has been identified as dubious, but it says nothing at all about why that hypothesis and not some other was the target of doubt and revision. To adequately solve Duhem's problem, i.e., to apply good sense, we need to be able to scrutinize this first-order judgment.

It is telling that although both Dorling and Howson and Urbach take themselves to be addressing the Duhem problem, they explicitly direct their work at the claims of Kuhn, Lakatos, Feyerabend, and their followers who, in Dorling's words, take "a cavalier attitude" toward refutation (Dorling 1979, pp. 186-187; Howson and Urbach 1993, pp.134-136). As we saw in the previous section, far from being "cavalier" about refutation, Duhem is making the very reasonable claim that beyond the reach of deductive logic, the scientist still has other decision-making tools.

But historical misreadings aside, since Duhem does not claim that first-order theoretical judgments are arbitrary, and in fact claims just the opposite, the Bayesian solution offered by Dorling and Howson and Urbach is really no solution at all. Ironically, by failing to provide any grounds for the reasonableness of choosing to revise one hypothesis, H, or another, these Bayesian accounts only serve to reinforce the "cavalier attitude" they were purported to refute. If we can say nothing more about how it was that the scientist determined the particular H in question to be the most suitable for revision, then we have failed to elucidate how the judgment was reached, much less justified, and have therefore failed to escape the charge of arbitrariness. The calculus they provide is suitable for refuting that claim that there can be no rational grounds for ever rejecting an hypothesis, but this is not Duhem's problem.

Despite the failure of Bayesian confirmation theory to provide an adequate treatment of scientific judgment in the face of underdetermination, it does still provide a unified framework to treat a very general class of decisions. The Bayesian accounts could show how, given two alternative hypotheses, one of them ought to be rejected in the face of recalcitrant evidence. Good sense must do at least this. As I have just argued, it must also do quite a bit more (i.e., the four activities enumerated above). The account I offer in the next section, leveraging work on scientific heuristics, provides the descriptive accuracy and depth needed to illuminate all four activities of Duhemian good sense.

1.3 Heuristics, Meta-heuristics, and Problem Space

The concept of a heuristic is not new to the philosophy of science. Whewell introduced the term, describing it as the "logic or art" of scientific discovery—in contrast to the logic of scientific deduction (cf. Todhunter 1876, p.418). Lakatos' methodology of scientific research programmes also afforded heuristics an important position, instructing researchers to avoid ("negative heuristic") or pursue ("positive heuristic") certain lines of inquiry (Lakatos 1965).

Tversky and Kahneman's landmark experimental work (1974) brings us close to the current understanding of heursitics, as fast and efficient problem-solving techniques that are especially useful when more thorough or exhaustive techniques are unavailable, impractical, or impossible. Their studies of scientific behavior suggest that heuristics actually do guide the judgments of scientists in practice. This fact, combined with Wimsatt's (1982, 2007) work on the necessity of using "reductionist research strategies"—families of simplifying heuristics—gets us off the ground. It is now uncontroversial that scientific judgments are grounded in heuristics, but the connection between this fact and Duhemian good sense is something new.

Wimsatt adopts his conception of heuristics, and their role in problem solving, from Simon's (1957) "principle of bounded rationality". Real human beings are not ideal rational agents. We do not, and in most cases *cannot*, employ complicated algorithmic procedures in order to reach a decision. And yet, we make good decisions all the time. Our success can be explained by heuristic strategies—using efficient, simplifying procedures, which, given our particular aims, transform a complex situation into one that can be more easily understood and judged.

Unlike an algorithm, a heuristic, when correctly applied to a problem, does not guarantee a correct solution. Nor does it guarantee any solution at all. Yet, these limitations may actually be *strengths*. As Wimsatt points out:

> The failures and errors produced using a heuristic are not random, but systematic. I conjecture that any heuristic, once we understand how it works, can be made to fail. That is, given the knowledge of the heuristic procedure, we can construct classes of problems for which it will always fail to produce an answer, or for which it will always produce the wrong answer (Wimsatt 1982, p.162).

The systematic errors produced by a heuristic are its bias. And the evidence of the bias in a particular instance, Wimsatt calls its “footprint” (Wimsatt 2007, ch.5). These properties open up a valuable line of methodological inquiry:

> Not only can we work forward from an understanding of a heuristic to predict its biases, but we can also work backwards, hypothetically, from the observation of systematic biases as data to conjecture as to the heuristic which produced them; and if we can get independent evidence as to the nature of the heuristics, we can propose a well-founded theory of the structure of our heuristic reasoning in these areas (Wimsatt 1982, p.162).

Wimsatt’s “well-founded theory of the structure of our heuristic reasoning” is a striking echo of the Duhemian project to elucidate the character of good sense. If scientific judgments are heuristics, then a “structural theory” of such judgments is very much in the spirit of Duhemian good sense. These are both projects to understand the kinds of problems for which any particular heuristic is well- or ill-suited and thereby provide methodological guidance for its fruitful application. Thus, I argue, that we can think of good sense as constituted by second-order, or meta-, heuristics—rough and efficient guidelines to the fruitful application of the first-order heuristics.

Recall also the four activities of good sense: (1) Judging the acceptability of fundamental assumptions; (2) interpreting experimental evidence and how it ought to impact upon theory, hypotheses, and auxiliary assumptions; (3) developing new theories and hypotheses; and (4) eventually resolving differences in scientific judgment. Leaving the fourth aside for the moment, it is remarkable that Wimsatt groups his examples of heuristics into categories that roughly correspond to these other three activities: (1′) conceptualization, (2′) observation and experimental design, (3′) model building and theory construction (Ibid., pp.174-175). The individual heuristics in these categories are prescriptions about what to assume, what to control in an experiment, what to simplify in a scientific model, etc. The meta-heuristics are the prescriptions about when we should adopt one heuristic over another.

Or to put it another way: If we think of heuristics as “actions” in a problem space; prescriptions to assume x, simplify the description of y, revise theoretical assumption A in light of evidence E, etc., then the meta-heuristics describe the structure of the problem space itself. Some meta-heuristics will be conceptual assumptions about *possibility*, delineating the problem or system of interest and the meaningful ways to go about exploring it. Others will be about *relevance*, specifying the appropriate variables, dimensions, or degrees of freedom. And still others will be about *prudence*, partitioning the space into three regions: (i) Those problems for which a particular heuristic is well-suited; (ii) those problems for which it is ill-suited (i.e., inappropriately biased); and (iii) those problems for which its suitability is unknown.

To illustrate how this all works, I will now analyze three examples: one from physics, one from emergency medicine, and one from ecological management. Although the language of heuristics, meta-heursitics, and problem spaces is not used by the researchers in these areas, I will show that their methodological analysis nevertheless fits that description.

1.3.1 Mathematical explanation in physics

The first example is from Wilson's (2007) and Batterman's (2009) discussion of modeling shock waves in physics. They invite us to imagine a long, gas-filled tube just as a short, violent pressure is applied to one end. We want to explain the behavior of the gas in response to the sudden pressure. Our fundamental particle physics tells us that the gas in tube is really a collection of molecules, so we might describe the situation thus:

> If a collection of the molecules are given a push (say by blowing into the tube at one end), then they will begin to catch up to those in front resulting in a more densely populated region separating two regions of relatively less molecular density. Across this region, molecules will exchange momentum with one another as if some kind of permeable membrane were present. The region occupied by this "membrane" is a shock (Batterman 2009, p.431).

Since tracking the movement of all the individual gas molecules is excessively complicated, to explain what happens in the tube, a heuristic, call it "ϕ_1", is applied:

ϕ_1 : Take the continuum limit.[5]

By applying ϕ_1, we idealize the gas inside the tube. Instead of treating it as we think it actually is—a dense collection of molecules—we treat it as a continuous fluid with little viscosity, shrinking the complicated "shock" down into a 2-dimensional boundary. This splits the gas inside the tube into two distinct regions, separated by the 2-dimensional boundary, whose behavior during the shock event can be treated with thermodynamical equations.

The heuristic, ϕ_1, is one member of a more general class of heuristics for *variable reduction* (Wilson 2007, pp.184-192). In cases of particularly complicated physical interactions, variable reduction heuristics prescribe that we divide up the description of the system into distinct regions, or "patches" in just this way. Each patch can then be dealt with by its own modeling equations, utilizing simplifications tailored to our explanatory needs. Batterman's description of the "modeler's methods of simplification" gives us further insight into how this is done:

[5] Throughout the thesis, I will adopt a convention for labeling first-order heuristics with lower case Greek letters and meta-heuristics with upper case Greek letters. In this chapter, I will use "ϕ" and "Φ" for physics, "θ" and "Θ" for emergency medicine (from the ancient Greek for "therapy"), and "o" and "O" for ecology (from the ancient Greek root of "ecology").

> First, one typically nondimensionalizes the equation or system of equations. This enables one to compare parameters appearing in the equation as to their importance or "size" even though they have been expressed in different units. Second, one takes limits thereby reducing the equation. Typically these limits involve letting a "small" nondimensionalized parameter approach the limiting value of zero or a "large" nondimensionalized parameter is taken to infinity. The aim is to simplify by idealizing in this fashion. . . . The hope is that if done correctly, one will end up with a model which exhibits the dominant features of the system. It will be a limiting model that displays the essential physics (Batterman 2009, pp.430-431).

A 2-dimensional boundary and thermodynamical equations are the end result of this process for the case of the shock. The details of the particles in tube—their positions, momentums, and interactions—are ignored for the purposes of explanation. Wilson acknowledges some of the "descriptive irony" of this situation: How is that we can explain what is going on in the tube by ignoring the complexities in the shock region?

> [T]he fact that a region can be descriptively avoided in this manner [i.e., idealizing the shock as a 2-dimensional boundary] does not indicate that it is therefore unimportant: the condition at the shock front represents the most important physical event that occurs in our tube. *It is merely that we can keep adequate track of its overall influence in a minimal descriptive shorthand* . . . Indeed, the whole idea of variable reduction or descriptive shorthand is that we are able to locate some shock-like receptacle that can absorb complexities and allow us to treat its neighboring regions in a simplified fashion (Wilson 2007, p.190, emphasis added).

A "minimal descriptive shorthand" is the hallmark of a simplifying heuristic. Applying ϕ_1 to treat the shock as a 2-dimensional boundary allows us to ignore irrelevant details about the system. We do not care about the initial configuration of molecules in each possible simulation of the event. What we care about is the way the shock moves and the way it effects the two regions on either side. These features can be explained in the manner Wilson and Batterman describe. The full, complicated story about how the individual molecules behave in the tube fails to adequately explain the phenomenon precisely because those details do not ultimately make a difference to the feature of the system we care about: the behavior of the shock event.

For Batterman, eliminating irrelevant details from a model is the right strategy when explaining universal patterns in nature. An explanation for some *particular* shock event may need to appeal to its initial molecular configuration; calculating all the complicated molecular interactions subsequent to the push. But this would not explain why we should expect the same (or similar) result next time, despite a completely different initial molecular configuration. To explain we why should, in general, expect the shock to behave as it does *requires* that we apply ϕ_1, appealing to the idealized, thermodynamic model.

Good sense, via Wilson and Batterman, thus gives a meta-heuristic governing certain kinds of mathematical explanations in physics:

Φ_1 : To explain universal behavior, use variable reduction heuristics.

The meta-heuristic Φ_1 is one of prudence. It prescribes a region of problem space for which variable reduction heuristics, like ϕ_1, are well-suited. But it is important to see that there are other regions, where Φ_1 and ϕ_1 do not apply. For example, suppose we want to explain the oscillations of a real pendulum. An ideal pendulum swings forever. Its behavior is dependent only upon the length of the rod, the amplitude of the swing, and its initial acceleration. Of course, no real pendulum swings forever, so in explaining why it is that a real pendulum will eventually stop moving, we must introduce additional variables into our ideal model. We can add a term to describe the friction at the pivot point. We can add a term for air resistance. Batterman discusses this approach as well:

> [T]he aim is to try and effect a kind of convergence between model and reality. Ultimately, the goal is to arrive at a complete (or true) description of the phenomenon of interest. Thus, on this view, a model is better the more details of the real phenomenon it is actually able to represent mathematically (Batterman 2009, p.429).

This gives us a second, common modeling heuristic:

ϕ_2 : Introduce more complexity and detail.

The heuristic, ϕ_2, can be thought of as belonging to a more general class of *model-world matching* heuristics, which Fowler (1997) describes thus:

> Applied mathematicians have a procedure, almost a philosophy, that they apply when building models. First, there is a phenomenon of interest that one wants to describe or, more importantly, explain. Observations of the phenomenon lead, sometimes after a great deal of effort, to a hypothetical mechanism that can explain the phenomenon. The purpose of a model is then to formulate a description of the mechanism in quantitative terms, and the analysis of the resulting model leads to results that can be tested against the observations. Ideally, the model also leads to predictions which, if verified, lend authenticity to the model. It is important to realize that all models are idealizations and are limited in their applicability. In fact, one usually aims to oversimplify; the idea is that if a model is basically right, then it can subsequently be made more complicated, but the analysis of it is facilitated by having treated a simpler version first (Fowler 1997, p.3, also quoted in Batterman 2009).

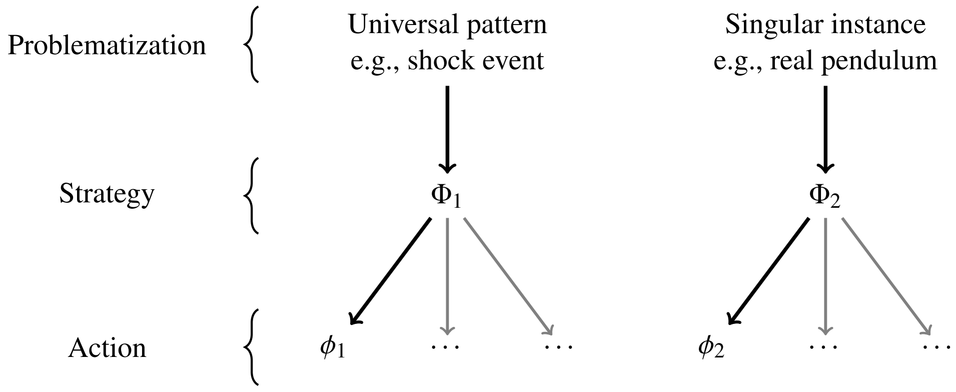

Figure 1.1: Problem space for mathematical explanation

The pendulum case exemplifies Fowler's strategy: An idealized model that can be made increasingly more complex as is needed. If we want to know why our real-world pendulum behaves as it does, then we need to make our model as much like the real system as possible, adding in the addition variables until we are satisfied with the accuracy of the model. Good sense, via Fowler, thus gives us a second meta-heuristic:

Φ_2 : To explain a singular, phenomenal instance, use model-world matching heuristics.

We can now contrast two meta-heuristics ranging over model construction for the purposes of mathematical explanation; descriptions of two different regions of problem space, as illustrated in figure 1.1. We begin with problematization and ask: "What do we want to explain?" If we seek an explanation for universal phenomena, then we follow the left path, where variable reduction is a necessary strategy in order to explain why the details of each individual system or event do not matter. For the case of the shock event, good sense, made explicit in Φ_1, prescribes that we apply the heuristic ϕ_1 to construct the model. On the other hand, if we want to explain a particular phenomenal instance, then we follow the right path. Convergence between model and reality is necessary; variables will need to be introduced until the match between model and reality is sufficient to our needs. For a case like the real pendulum, Φ_2 is our strategy, and prescribes that we apply ϕ_2.

This "map" of the problem space makes explicit how meta-heuristics connect our scientific aims, like explanations, to specific "actions", like applying a modeling heuristic. Obviously, the story and problem space representation I have provided here is incomplete. One heuristic is not enough to construct a model, nor are these two kinds of explanations and strategies exhaustive of all possible strategies. In further elucidating modeling heuristics and the meta-heuristics of good sense that prescribe their use, we should want to say much more about the different kinds of problems dealt with by mathematical models in physics. This is a project unto itself, and since it is not the focus of my work, I shall not pursue it further here. This brief illustration is only meant to show how the framework of heuristics and meta-heuristics I propose can accommodate one kind of methodological divergence; answering the Duhemian problem of why one physicist might choose to take the continuum limit rather than introducing additional variables into her model (or vice-versa).

Finally, I should note that figure 1.1. is just one way to go about representing the problem space of mathematical explanations. I have organized the process according to three discrete methodological stages; answering three different questions: What kind of explanation do we seek? What is the general modeling strategy that applies? How do we construct the model? Alternatively, one could also use the phenomenon to "shape" the overall space, for example, starting from the observed behavior of shock events and then organizing the meta-heuristics according to the kind of explanatory question they answer. Or one could arrange the space as a kind of continuum, observing that applications of ϕ_1 and ϕ_2 are inversely related. Modelers must begin somewhere along the continuum, striking a certain balance between variable reduction and model-world matching. As they continue to apply ϕ_1 or ϕ_2 they slide along this modeling continuum. Thus, the philosophical point is not the one representation of problem space I described here, but the *process* of organizing the problem space itself. What representation of the relevant heuristics will be most fruitful will depend upon the aims of the investigation.

1.3.2 Decision-making in emergency medicine

This second example comes from Gigerenzer and Gaissmaier (2011), who desribe the problem thus:

> When patients arrive at the hospital with severe chest pain, emergency physicians have to decide quickly whether they suffer from acute ischemic heart disease and should be assigned to the intensive coronary care unit (ICU). In a Michigan hospital, doctors preferred to err on what they believed was the safe side by sending about 90% of the patients to the ICU, although only about 25% of these actually had a myocardial infarction (Green and Mehr 1997). The result was an overly crowded ICU, a decrease in quality of care, an increase in cost, and a risk of serious infection among those who were incorrectly assigned. Something had to be done (Gigerenzer and Gaissmaier 2011, pp.467-468).

Green and Mehr (1997) designed a trial to test the clinical value of introducing a decision-support tool, the Heart Disease Predictive Instrument (HDPI), a logistic regression model requiring the physicians to assess and calculate around 50 different probabilities depending upon the presence or absence of various symptoms. Their finding was somewhat surprising:

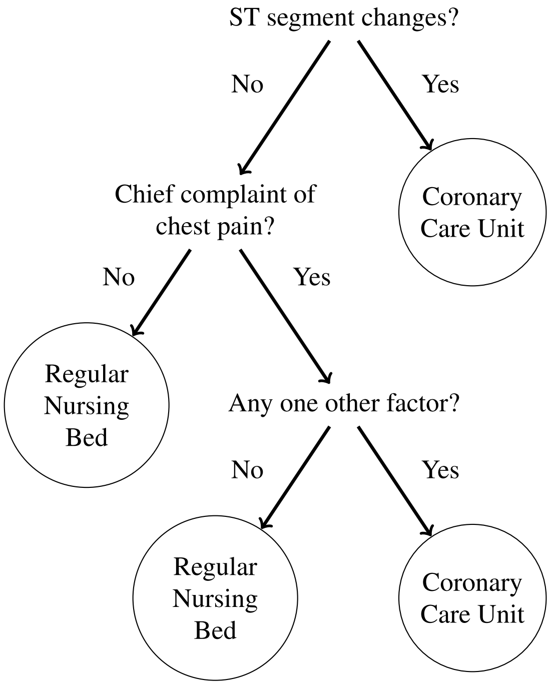

Figure 1.2: Fast-and-frugal tree for medical decision-making

> Using the HDPI as a decision-support tool, we attempted to improve [ICU] utilization from historically high levels, only to find that utilization had already changed. The change occurred after the resident physicians learned about the HDPI but before they began using it to actually calculate probabilities. The change persisted for months after the intervention was withdrawn (Green and Mehr 1997).

Green and Mehr note that their result is consistent with a generally “disappointing” trend of physicians simply ignoring similar decision-tools after their introduction. However, they suggest that the timing of the observed improvement could be explained by the physicians using the knowledge of the relevant diagnostic factors indicated on the HDPI in order to construct simpler heuristics, like the “fast-and-frugal” decision tree shown in figure 1.2. The heuristic decision tree has no probabilities, requires no calculations, and since it contains only three yes-or-no questions, it is easy to memorize and apply. It was also shown to be more accurate than the HDPI at predicting actual heart attacks (Gigerenzer and Gaissmaier 2011, p.468).

The decision tree of figure 1.2 is a model that results from applying a simplifying heuristic to the problem of how to assign possible acute ischemic heart disease (AIHD) patients to the ICU. Let us summarize this heuristic like so:

θ_1 : Use a fast-and-frugal decision tree.

Applying θ_1 reduces the complexity of the problem down to three yes-or-no questions about a few diagnostic indicators (e.g., “ST segment changes” on the electrocardiogram, complaints of

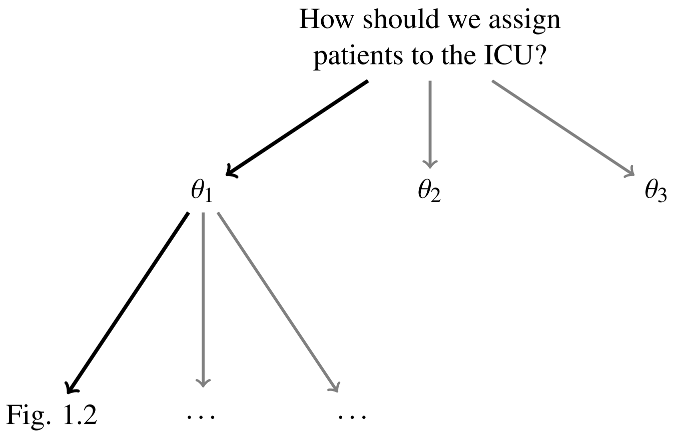

Figure 1.3: Problem space for improving ICU assignment

chest pain). But as this case shows, it is not the only way to simplify the problem. Green and Mehr's study was a test of a different heuristic:

θ_2 : Use a logistic regression model.

Applying θ_2 results in the HDPI, which was seemingly ignored by the physicians, who, before the study began, were applying the much cruder heuristic:

θ_3 : Err on the side of caution.

The result of applying θ_3 was an ICU overfull with patients who did not really need to be there, taking up resources from patients who actually did need immediate care.

Figure 1.3 provides a representation of this problem space. We have three different heuristics, all applicable to the problem of ICU assignment. Gigerenzer and Gaissmaier share Green and Mahr's conclusion that θ_1 is the best way to go, since it provides a rule that physicians can actually use (making it better than θ_2), is more efficient than θ_3, and to top it off, is the most predictively accurate.

But there is no reason to think that the particular decision tree in figure 1.2 is the only model that results from applying θ_1. The diagnostic factors used there might not turn out to be the best predictors. Or it might be possible to reduce the questions down to two, further increasing efficiency without any loss in accuracy. θ_1 thus prescribes a kind of action—to use a fast-and-frugal tree—but it does not tell us the content of that tree. For this, additional judgments, guided by additional evidence and heuristics, will be needed. Much the same can be said for θ_2. The HDPI is not the only possible logistic regression tool.

It is also important to see that Green and Mehr's study was not ill-conceived. It is not an obvious fact that the HDPI would be ignored by physicians and thus rendered irrelevant to

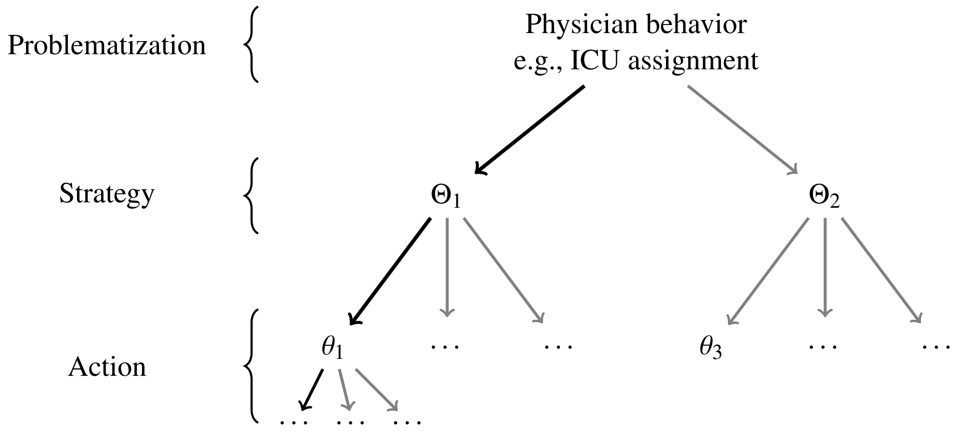

Figure 1.4: Problem space for improving physician behavior

improving ICU efficiency and assignment accuracy. Despite the fact that it was not the stated intention, their study nevertheless provides us with useful evidence about meta-heuristics, relevant to the larger problem space of improving physician behavior. Their conclusion to recommend further work with fast-and-frugal decision trees can be represented with a meta-heuristic like so:

Θ_1 : To improve physician behavior, apply θ_1.

This is in contrast to the hypothesis they tested, which we can think of as guided by an alternative meta-heuristic:

Θ_2 : To improve physician behavior, apply θ_2.

Figure 1.4 represents how these two meta-heuristics can be arranged in a more general problem space, concerned with viable strategies to improve physician behavior. It is analogous to the problem space of mathematical explanation in figure 1.1, wherein we want to compare meta-heuristic strategies and the resulting heuristic prescriptions. But in contrast to the case from physics, for which the paths of meta-heuristics Φ_1 and Φ_2 were appropriate to specific kinds of explanations, the lesson from this case is that one of these two strategies provides an inferior solution. Θ_1 is the exclusive strategy of good sense here, prescribing that θ_1 is well-suited to the problem of improving ICU assignment, while θ_2 is not.

There are two more points I want to draw out from this case. First, it elucidates the ways in which our understanding of problem space can be enriched by empirical work. At the time of Green and Mehr's study, Θ_1 was (presumably) unknown. Θ_2, with its prescription to implement a tool like the HDPI, appeared to be the only available strategy. The unusual way in which Θ_2 failed revealed a fact about the structure of the problem space: There was another strategy at work!

Second, knowledge of the overall problem space is not necessarily relevant to all of the scientists acting within the space. The emergency physicians, for example, need not be concerned about any regions above the level of action. They are focused on assigning patients the appropriate care, so all they want is the right heuristic for th task, whatever that may be. Whereas the clinical researcher (or social psychologist) may be interested in the different possible heuristics that physicians could use, each with its respective merits and biases. For them, the regions of meta-heuristic strategies and problematization are important to understand. For the philosopher of science, and the project in this thesis, the interest is precisely in presenting the overall picture; to illustrate the connections between the aim of inquiry, which guides the heuristic reasoning, which in turn guides the final scientific judgment.

1.3.3 Resilience in ecological management

The final example on this chapter involves a growing dissatisfaction in ecology with the traditional "optimization approach" to ecological modeling and management. Collapse of commercial fisheries, chronic pest outbreaks in spite of rigorous pest control measures, and failures of flood control and drought prevention are but a few examples of the kinds of system management failures that have contributed to the dissatisfaction (cf. Gunderson and Holling 2002). Walker and Salt (2006) characterize the optimization approach thus:

> An optimization approach aims to get a system into some particular "optimal state", and then hold it there. That state, it is believed, will deliver maximum sustained benefit. It is sometimes recognized that the optimal state may vary under different conditions, and the approach is then to find the optimal path for the state of the system. This approach is sometimes referred to as a maximum sustainable yield or optimal sustainable yield paradigm.
>
> To achieve this outcome, management builds models that generally assume (among other unrecognized assumptions) that changes will be incremental and linear (cause-and-effect changes). These models mostly ignore the implications of what might be happening at higher scales and frequently fail to take full account of changes at lower scales.

In addition to size scales, this modeling approach tends to overlook time-scales. It optimizes for the short-term, seeking to achieve a rigid system stability.

The assumptions of the optimization approach that Walker and Salt describe can be stated as different modeling heuristics:

o_1 : Assume ecosystems have an optimal state.

o_2 : Manage ecosystems by holding them in the optimal state.

o_3 : Assume changes to ecosystem states are incremental and linear.

o_4 : Treat the system at a single time and size scale.

While this list of heuristics is obviously not exhaustive, for the purposes here, we can think of the optimization approach as a meta-heuristic prescription to use this set:

O_1 : For ecological management models, use $o_1 \ldots o_4$.

Applying O_1 results in an idealized model of ecosystem behavior. This is not a serious problem in itself, since all scientific models are idealized in some way. The problem only arises because natural ecosystem behavior is so different from these models. Natural ecosystems are not governed exclusively by short-term forces operating at a single size scale. Shocks to an ecological system can come from any scale, and if a system is managed too rigidly, unable to absorb these shocks, it is liable to suffer a "system flip", a non-linear change in its dynamics, dominated by unfamiliar process and (typically) unwanted effects. Pest outbreaks, floods, and droughts are all examples of such system shocks, potentially leading to system flips.

Attempts to manage outbreaks of the spruce budworm in spruce/fir forests in North America provide a vivid (and now well-worked) example of O_1's imprudence and the pernicious biases introduced by its heuristics. Budworm outbreaks were observed to occur every 40 to 120 years, and could kill up to 80% of the spruce/firs. This was economically damaging to the forest industry, so they began a massive pesticide campaign to control the population of budworms, thinking they could achieve a stable, low budworm population while optimizing the yield of their desired resource. Walker and Salt point to the problem:

> In a young forest, leaf/needle density is low, and though budworms are eating leaves and growing in numbers, their predators (birds and other insects) are easily able to find them and keep them in check. As the forest matures and leaf density increases the budworms are harder to find and the predators' search efficiency drops until it eventually passes a threshold where the budworms break free of predator control, and an outbreak occurs.
>
> While the moderate spraying regime avoided outbreaks of budworms, it allowed the whole forest to mature until all of it was in an outbreak mode. Outbreaks over a much greater area were only held in check by constant spraying (which was both expensive and spread the problem). . . . Now there was a critical mass of tree foliage and budworms. The whole system was primed for a catastrophic explosion in pest numbers. The managers in this system were becoming locked into using ever-increasing amounts of pesticide because the industry wouldn't be able to cope with the shock of a massive pest outbreak (Walker and Salt 2006, p.96).

Analyzing this and similar cases inspired C. S. Holling's (1973) proposal of a "resilience approach" to ecological management. He argues that the optimization goal of achieving a stable equilibrium state is unobtainable; will tend toward these failing policies; and increase the likelihood of an ecosystems flip into a degraded state. A resilience approach aims to achieve a different kind of stability over time, one that expects an ecosystem to pass through what he calls the "adaptive cycle" (Gunderson and Holling 2002).

The adaptive cycle consists of four (typically consecutive) phases:

1. Growth phase, in which new opportunities and available resources are exploited;

2. Conservation phase, in which resources are stored with increasing efficiency and interconnection;

3. Release phase, in which a shock disrupts the interconnections in the system and the stored resources are released;

4. Reorganization phase, in which new opportunities are created.

The adaptive cycle is a complex, conceptual heuristic, call it o_5 (i.e., "use the adaptive cycle to model ecosystem behavior"). It prescribes that a system be analyzed and modeled according to certain simplifications—specifically, that its dynamics can be organized into four discrete phases and that each phase is meaningful for the system in question. It is similar to θ_1 from the last example in the sense that it instructs us to build and apply a kind of model, but does not provide the content of that model. θ_1 tells us to construct a fast-and-frugal tree, but does not tell us what are the particular decision forks. o_4 tell us to break down the system dynamics into these four phases, but does not tell us what each phase will look like for any given system.

For some systems, the content of the adaptive cycle is easy to see. In the budworm case, each outbreak is a release phase. When the trees are killed, they free up space and resources for a reorganization phase. Young trees seize on these newly available resources in a growth phase. As the forest matures in the conservation phase, foliage density increases and it becomes primed for another release phase.

But like all heuristics, the four-phases of the adaptive cycle do not apply to every system. For example, a bog's slow transformation into a forest does not pass through a release phase, and a young forest may never accumulate and store enough resources to enter a conservation phase (Walker et al. 2006, p.4). Recognizing the conditions for the adaptive cycle's applicability is thus key to effectively employing a resilience approach.

A resilience approach also includes at least two other heuristics:

o_6 : Do not resist the phase changes of the adaptive cycle.

o_7 : Manage ecosystems to prevent system flips.

In the spruce/fir forest, this means that managers should not attempt to prevent any release phase. Rather, they should allow outbreaks to proceed, aiming instead to contain the outbreaks and prevent a forest-wide pest explosion. The massive insecticide campaign against the budworm sought to hold the entire forest in the conservation phase. But in doing so, it brought the forest into an ever more fragile conservation phase. Once Holling recognized and articulated the adaptive cycle, a new policy of spraying was implemented, one which sprayed less frequently and only in specific regions of the forest. Outbreaks could then be contained while allowing the forest to cycle through its phases (Walker and Salt 2006, p.80).

We can summarize the resilience approach with an alternative meta-heuristic:

O_2 : For ecological management models, use $o_5 \ldots o_7$.

Like O_1, the management strategy of O_2 will produce idealized models. However, its model of ecosystem behavior will be far less static, and more closely resemble the behavior of actual systems. The strategy of O_1 reflects what Gunderson and Holling call the "myth of nature balanced":

> Hence if nature is disturbed, it will return to an equilibrium through (in systems terms) negative feedback. Nature appears to be infinitely forgiving. It is the myth of maximum sustainable yield and of achieving fixed carrying capacities for animals and humanity. It imposes a static goal on a dynamic system (Gunderson and Holling 2002, p.12).

This is in contrast to O_2, which adopts a different picture:

> The [resilience approach] is a view of multistable states, some of which become irreversible traps, while others become natural alternating states that are experienced as part of the internal dynamics. Those dynamics result from cycles organized by fundamentally discontinuous events and nonlinear processes. There are periods of exponential change, periods of growing stasis and brittleness, periods of readjustment or collapse, and periods of reorganization and renewal. Instabilities organize the behaviors as much as stabilities do (Ibid., p.13).

The failure of the optimization approach to successfully manage ecosystems, like the spruce/fir forest, is a recalcitrant experimental result. Holling's insight was to see that the failure was not limited to the solution of overspraying, but to the underlying strategy, made explicit by the heuristics $o_1 \ldots o_4$. The resilience approach is therefore a deep theoretical revision, reconceiving how we ought to think about ecosystem dynamics.

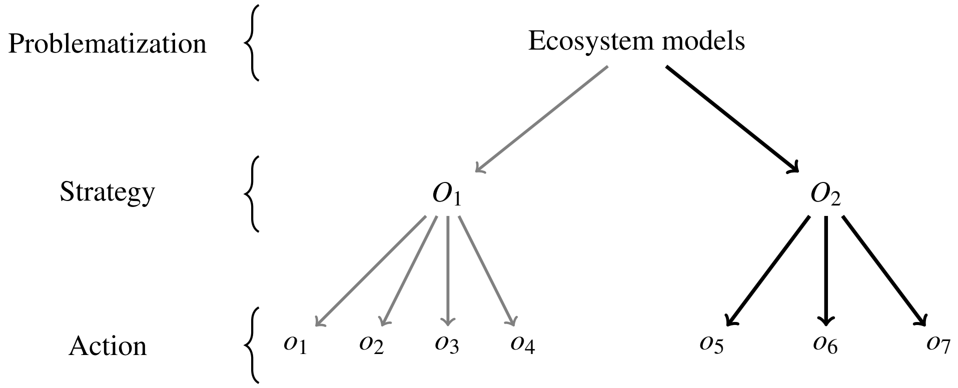

Figure 1.5: Problem space for ecological management

The relevant problem space is summarized in figure 1.5. Like the emergency medicine example, the lesson of good sense here is to come down in favor of O_2. For the problem of ecological management, the heuristics prescribed by O_1 produced unusably biased models, leading to poor management policies.

Chapter Summary

The arguments of this chapter have all been to suggest the power of a heuristic framework for expanding upon Duhemian good sense; showing how it can be extended and applied to understand certain problems in the methodology of science. Each of these three examples elucidates an important methodological contrast. For physics, we saw a distinction between modeling approaches; in medicine, a distinction between decision-making tools; and in ecological management, a distinction in management strategies. As different as these sciences may be, the fact that they are all susceptible to my proposed framework and analyses of good sense speaks to the general applicability of my approach.

The philosophical strategy throughout much of the thesis will proceed as it has here: Beginning with an extant methodological controversy or conflict, I will look to unpack the relevant heuristics at work, tie these to meta-heuristic strategies, and then construct a plausible picture of the problem space in order to resolve or clarify the nature of the conflict. While the examples of this chapter have served to demonstrate the breadth of my approach, the remaining chapters will serve to demonstrate its depth.

In the next chapter, I will focus on a particular class of meta-heuristics dealing with stability and robustness. Although these terms are often used synonymously, I argue that there is a principled distinction to make between the study of stability, to be understood as a branch of mathematics, and robustness, to be understood in two ways: (1) as a *property* of results invariant under perturbation to heuristics, and (2) as a meta-heuristic *strategy* for that heuristic perturbation.

Despite its importance, the concept of robustness has not been free from philosophical critique. Critics have charged that the virtues of robustness are too easily undermined by the potential for pseudo-robustness and the difficulty in handling evidential relevance or discordance. Although each of these criticisms is not without merit, I argue that once the concept of robustness is clarified, as I have suggested above, then the criticisms cease to be troubling.

In chapter 3, armed with a clarified understanding of robustness, I will deploy my framework of good sense toward an analysis of the group selection controversy from the history and philosophy of biology. The methodological lessons emerging from this well-worked controversy are diverse and somewhat scattered. My aim will therefore be to bring many of the insights together, showing how alternative strategies in studying and interpreting the results about group selection can be unified to articulate a meta-heuristically rich problem space, useful for both retrospective analysis and prospective research design.

In chapter 4, I apply the framework to a contemporary methodological debate in clinical research, where the orthodox view is that every clinical trial needs to have *assay sensitivity*—the prospective ability of a trial to distinguish between an effective and ineffective treatment. Assay sensitivity can only be assured, it is claimed, with the use of a placebo control. However, running a placebo-controlled trial when there is already a known effective treatment deprives study participants of competent medical care. Thus, a conflict arises between doing valid science and doing ethical science.

I argue that this orthodox view is mistaken and rests upon no less than four errors: (i) conflating biological efficacy and clinical effectiveness; (ii) oversimplifying the epistemology and ethics of placebos; (iii) assuming an implausibly radical form of underdetermination; and (iv) conflating the analysis of a *trial-as-designed* with a *trial-as-executed*. Once these errors are corrected, the conflict between good science and ethical science dissolves.

Following that, I bring the insights of heuristics and robustness from the previous two chapters to bear on the science of clinical trials, laying out a general model of problem space for the kind of grand, scientific project that is a clinical research program. Then, in chapter 5, I test this model by applying it to the current state of research with the anti-tuberculosis agent, *moxifloxacin*. Results from the various experiments with moxifloxacin are discordant, resisting the easy, heuristic solutions of just doing more experiments or performing a statistical meta-analysis. I show how the model can be used to more precisely characterize the failure of robustness at the point of translation between mouse models and human studies.

I conclude in chapter 6 by revisiting the main themes and conclusions, as well as, discussing some possibilities for future extensions of the ideas in this thesis.

Chapter 2

Robustness and Its Critics

The epistemic necessity and practice of using simplifying heuristics creates a prima facie conflict with the search for truth and reliability in science. Since heuristics bias our theories, models, and experiments in certain ways, how can we trust the predictions they give us? How do we know when a result is not just an artifact of the simplification? This kind of question is made all the more pressing when we consider heuristics like variable reduction, which demand that we leave out details. How do we test to ensure that the small set of variables we focus on are the right ones? In sum: How do we get reliable results out of what are essentially false assumptions and biased models?

To answer these sorts of worries in biology, Levins (1966) proposes a particular conception of *robustness*. For Levins, robustness is a property of results that are shown to be free from the artifacts and biases of underlying assumptions. He writes:

> There is always room for doubt as to whether a result depends on the essentials of a model or on the details of the simplifying assumptions. ...Therefore, we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their different assumptions, lead to similar results we have what we can call a robust theorem which is relatively free of the details of the model (Levins 1966, p.423).

Since it is neither useful nor possible to construct a perfect one-to-one model of any system, Levins' argues that some simplifications must be made. He proposes five dimensions of modeling—generality, realism, precision, manageability, and understandability—some of which must be sacrificed or reduced for the sake of increasing others.[1] In doing so, modelers will make simplifying assumptions, and these assumptions will bias their models, making them

[1] Foreshadowing later discussion, Levins' five dimensions can be understood as the result of a meta-heuristic judgment. He is delimiting the characteristics of a model that are relevant for understanding the model's contribution toward a robust exploration of problem space.

false in the strictest sense. But this is no great problem. Multiple false models, which agree on a result despite their different assumptions, can still give us good reason for thinking that the result is true. As he puts it, "our truth is the intersection of independent lies" (Ibid.).

This concept of robustness, as "multiple means of determination to 'triangulate' on the existence and character of a common phenomenon, object, or result" (Wimsatt 2007, p.43) can be traced back at least as far as Whewell's idea of the *consilience of inductions*:

> The Consilience of Inductions takes place when an Induction obtained from one class of facts, coincides with an Induction, obtained from another different class. This Consilience is a test of the truth of the Theory in which it occurs (Whewell 1840, p.xxxix).

As Wimsatt points out, ideas similar to Whewell's can be found even earlier—in the scientific method of Aristotle and in the distinction between primary and secondary qualities in Galileo (Wimsatt 2007, p.43). Nevertheless, it is Levins who lays out the modern conception of robustness with its relationship to modeling. And it is Wimsatt (1981) who most emphasizes its connection to reliability:

> [A]ll the variants and uses of robustness have a common theme in the distinguishing of the real from the illusory; the reliable from the unreliable; the objective from the subjective; the object of focus from artifacts of perspective; and, in general, that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting (Wimsatt 1981, p.128).

But despite the obvious importance suggested by this passage, just what robustness is, and how it fits into a methodology of science remains somewhat obscure. Levins, for example, writes that "[t]he search for robust theorems reflects the strategy of determining how much we can get away with not knowing, and still understand the system" (Levins 1993, p.554). Robustness is a property of "theorems", but it also "reflects" a certain epistemic strategy. This latter aspect of the concept—as part of a strategy—is lost in Orzack and Sober (1993) who take it to mean only that "a statement's *robustness*, as distinct from its *observational confirmation*, can be evidence for its truth" (Orzack and Sober 1993, p.538, original emphasis) or Stegenga (2009), who defines it as "[t]he state in which a hypothesis is supported by evidence from multiple techniques with independent background assumptions" (Stegenga 2009, p.651).

Wimsatt (2007) and Weisberg (2006), on the other hand, focus on "robustness analysis" as a *procedure* for evaluating invariance conditions. Their views are closer to Levins, since robustness is both a property and a procedure (i.e., "strategy") for finding that property.

No one denies that the search for robust results is ubiquitous in science. However, critics of the concept have focused on the robustness property, pointing out difficulties for its fruitful or explanatory use. In this chapter, I will argue that these criticisms fail for ignoring the relationship between robustness as a property of results and robustness analysis as a strategy to achieve reliable answers. Building on the framework for good sense in the last chapter, I argue that we can think of the strategy of robustness with the following meta-heuristic:

> Γ_1: To minimize the possibility of bias and error, perturb heuristics across the relevant problem space of model/experimental design.

This is a research meta-heuristic, prescribing how to build and modify models or experiments to achieve the desired end of a robust and reliable result. But before we can explore its implications, there are some fundamental terminological issues to sort out. I will begin in the next section by showing how the language of robustness is often conflated or used interchangeably with the language of mathematical stability. While this is not necessarily a serious problem, the domains of mathematical stability and robustness (in Levins' and my sense) are not the same, and therefore, it is worth distinguishing clearly between these.

In §2.2, I will discuss an objection to robustness stemming from Levins' requirement of "multiple, *independent* models". Unfortunately, the sense of independence he is after is unclear. Orzack and Sober (1993) rightly point out that neither logical nor statistical independence will suffice. While independence in any strict sense is unobtainable, I will argue that what is necessary for robustness is instead a *diversity* of heuristics across the sampled models. I will then show one way to illustrate model diversity, extending a methodological tool from Thorpe et al.'s (2009) analysis of clinical trial design.

Following that, in §2.3, I will address two related concerns articulated by Stegenga (2009):

i The problem of discordant data: When different kinds of models or experiments produce inconsistent or incongruent results, how can these be brought to bear upon considerations of robustness?

ii The problem of relevance: Given a large set of discordant data, how should scientists distinguish the high quality and relevant evidence from the low quality and irrelevant evidence?

I argue that the meta-heuristic understanding of robustness as a strategy renders these problems benign. There is no reason to think that perturbing heuristics to find robust results should be straightforward. Therefore, while both of Stegenga's objections are important considerations for a robustness analysis, they do not undermine its value as a meta-heuristic strategy.

The ultimate aim of this chapter is to establish a central place for robustness considerations, as both a property of results and a research strategy, within my framework of good sense.

Heuristics guide judgments, producing the models, experimental designs, and epistemic tools that scientists use, but these are all in the pursuit of reliability. A result that is a mere artifact of scientific representation is ultimately not of much value. A robust result, trustworthy and free from bias, is the goal of inquiry.

2.1 The Language of Robustness

An understanding of robustness as the property revealed by "shared results" from multiple models or kinds of investigations is not universal in science or philosophy. D'Arms et al. (1998), for example, refer to a *robust* result in evolutionary game theory as an end-state that is indifferent to initial conditions. This is in contrast to a *stable* result, an end-state that is resistant to "invasion by rival strategies", i.e., once an equilibrium state of the system has been reached, the more stable it is, the more difficult it will be to perturb the system from that state (D'Arms et al. 1998, p.82).

Jen (2005) offers a different distinction between stability and robustness. For her, *stable* refers particularly to stability theory in mathematics. A system is stable if "small perturbations to the solution result in a new solution that stays 'close' to the original solution for all time" and *structurally stable* if "small perturbations to the system itself result in a new dynamical system with qualitatively the same dynamics" (Jen 2005, p.8). On her view, robustness is something broader than these, characterizing invariance relationships across a more varied class of systems, perturbations, and features that are beyond the scope of stability theory (Ibid., p.9).

Complicating matters further, Weisberg and Reisman (2008) refer to Jen's two kinds of "stabilities" as *parameter robustness* and *structural robustness*, respectively. Then they add the third category of *representational robustness*, derived from Levins' conception: Since no model is a perfect mirror of reality, then one can increase the reliability of a model's result by verifying it with multiple other models, each representing the system in a fundamentally different way.

"Robustness", "stability", and "parameter stability" are thus used by three different authors to refer to essentially the same property: the invariance of a result under perturbation to initial conditions.[2] Similarly, "structural stability" and "structural robustness" both get used to refer to the invariance of a result under perturbation to parameter values or equations. While this overlap of terms is not necessarily problematic, it is clearly less than optimal.

[2] Strictly speaking, Weisberg and Reisman's concept of parameter stability is misleading, since initial conditions are variables, not parameters. A variable is the argument of a function; the value that changes as one considers different initial conditions. A parameter characterizes the function itself, and thus, perturbations of parameters should be understood as falling into the category of structural stability.

To tidy this up a bit, I propose a qualitative division between the different kinds of mathematical stability, on the one hand, and Levins' concept and "representational robustness", on the other. Stability is a proper subject of mathematics, dealing with dynamical systems and their sensitivity to perturbations of various kinds. It is possible to provide rigorous proofs and demonstrations of results that are *Lyapunov stable*, when they are invariant across initial conditions (cf. Lyapunov 1966); and *structurally stable*, when they are invariant across parameter values and equations (cf. Smale 1980). A stability proof shows us something about the relationships among variables and parameters internal to a mathematical model or family of mathematical models.

Importantly, stability does not directly inform us about the natural world. Mathematical models can and do yield reliable predictions about the natural world, but the stability of their results is not a sufficient justification for trusting their predictions about natural phenomenon. A mathematical result could be perfectly stable without representing anything in the natural world.

The translation of results from the model to the world is where robustness enters the picture. Indeed, the threat of bias or error in this translation was Levins' motivation for introducing the concept. Here we can draw the clear, qualitative line: Stability is an *invariance* consideration *internal* to a mathematical model. Robustness is ultimately a *reliability* consideration *external* to a (not necessarily mathematical) model. But robustness also requires a kind of invariance. For Levins, showing that a result is robust means showing that it is invariant across "simplifying assumptions". For example, he takes the proposition that "in an uncertain environment, species will evolve broad niches and tend toward polymorphism" to be robust, and shows how it is invariant across three different models: a fitness set, a calculus of variation argument, and a genetic system. These models all share some common subject matter, since each must have a way to represent an uncertain environment and its effects on polymorphism, yet there is no mathematical transformation to turn one model into the other. The fitness set assumes discrete, different environments that can have relational effects on phenotypes, whereas the calculus of variation argument allows for a continuum of environments in which fitness in one environment contributes nothing to fitness in another. Neither of these two models address the underlying genetic system (Levins 1966, pp.423-427).

The qualitative difference in simplifications amongst his three models may be intuitively clear, but we can take the analysis one step further, and recognize that a simplifying assumption is the result of applying a heuristic. Moreover, a single heuristic can give rise to superficially different simplifications that nevertheless all contain the same bias.[3] This is a serious problem,

[3] The case-study on group selection in biology discussed in the next chapter provides an illustration of this problem: Seemingly different assumptions all arising from an application of the same modeling heuristics, giving

since shared bias is exactly the problem robustness is supposed to solve.

To resolve this problem, we can modify our understanding of robustness to mean *invariance across different heuristics*—a shift in emphasis that elucidates an important interdependence of robustness and heuristics. Despite Levins' pithy remark that "our truth is the intersection of independent lies", it is not obvious how multiple, false assumptions, albeit false in different ways, contribute to a result's reliability. However, if we recognize that Levins' "false assumptions" are the result of applying heuristics, then a plausible answer becomes available. Heuristics are necessary to simplify the description and treatment of a problem. But in doing so, they introduce bias. Robustness is a way to control for bias. By perturbing the heuristics across the space of sampled models, we can show that a result is free from any single kind of bias. Moreoever, heuristics are unlike their resulting assumptions and simplifications in that they can have some degree of "built-in" reliability. They are neither true nor false, but epistemic tools adapted to a specific purpose. Insofar as the relevant meta-heuristics and problem space are well-understood, i.e., elucidating how the heuristic is appropriate for the problem domain, then we can be confident that the shared result is genuine.

Returning to Levins' example, we can illustrate this re-conception of robustness. The problem he considers is the effect of an uncertain environment on species polymorphism. Across his three models, there are two alternative heuristics of conceptualization regarding the level of organization:

γ_1: Treat the problem at the organizational level of the organism.

γ_2: Treat the problem at the organizational level of the gene.

The simplifications of our model will change dramatically depending upon which conceptualizing heuristic, γ_1 or γ_2, we apply. Insofar as Levins' result is invariant across these two heuristics, then that is positive evidence of his results' robustness. But we can also see that all three of his models, by only exploring a single level of organization, will be biased against interactions or dependencies that may overlap these levels. This appreciation for the kinds of heuristics that have not yet been incorporated by the sampled models, or the recognition of a similarity in the heuristics' bias, speaks to *robustness as a strategy* and the need for meta-heuristic, Γ_1. We might wonder whether or not Levins' result is robust across the right set, or a sufficiently diverse set, of heuristics (more on this in the next section). If we are not interested in interactions between levels of organization, then we can reasonably judge Levins' set to be adequately robust.

But notice that there is no obvious metric by which we can measure the difference between heuristics. We can easily recognize that Levins' three models are not identical, trace their

rise to a pseudo-robust result.

structure back to the kinds of heuristics used for their construction, and then see that these too are not identical. This provides the minimal epistemic grounds for asserting the property of robustness—the heuristics used to construct these models are different, yet they agree on a result. Thus, we have good reason to think that the result is not an artifact. But to say more than this; to talk about the *extent* to which the result may be robust, we will need to begin thinking about the relevant problem space.

Despite the distinction I want to draw, this last point, about the extent of robustness, shows that robustness and stability do still have some properties in common. Each can come in degrees and the existence of one does not guarantee the existence of the other. Most importantly, as Jen points out, "[i]t makes no sense to speak of a system being either stable or robust without first specifying both the feature and the perturbations of interest" (Jen 2005, pp.8). In other words, when talking about either stability or robustness, we always need to ask "stability or robustness of what to what?"

For stability, the answers to this question can often be precisely stated. A feature or result is Lyapunov stable when it is invariant (usually within some defined area of state space) over changes to the initial conditions. A feature is structurally stable when it is invariant over changes in the parameter values or governing equations. Both of these speak only about mathematical dependencies, where it is possible to describe the exact thresholds (e.g., values of key variables and parameters) where the stability holds or fails. But importantly, they do not, without further justification, speak toward any kind of reliability in the natural world. The inferential jump from mathematical stability to reliability in the natural world requires some further assumptions about the applicability of the mathematical model in question.

Answering the "*robustness* of what to what" question gets us closer to reliability, but is going to be more complicated. I have argued here that a result is robust when it is invariant across different heuristics. However, since heuristics and their resulting simplifications can hide biases, achieving robustness calls for a well-explored and well-understood problem space, rich in meta-heuristics to elucidate the relevant models, biases, and aims of research.

Finally, my picture of this strategic and terminological problem space is summarized in figure 2.1. To explain the natural world, we need a strategy of heuristic perturbation. To explain the internal workings of a mathematical model, we need a strategy of stability of analysis, perturbing the initial conditions and parameters. The arrow between the two kinds of explanations reflects an analogical relationship between our understanding of a model and our understanding of the world. For models whose potential biases are recognized and determined to be benign for the problem under consideration and whose predictions have been accurate, an understanding of their internal relationships may be an appropriate surrogate for the relationships in the natural world. In other words, if the meta-heuristics of a problem space are well-understood,

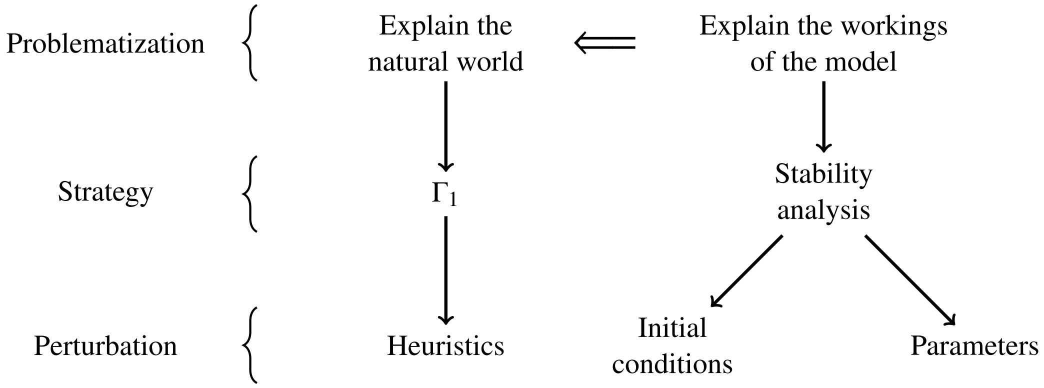

Figure 2.1: Problem space for robustness and stability considerations in modeling

then the internal features of a model (or family of models) may provide sufficient justification for a robustness claim.[4]

2.2 Independence or Diversity

One such meta-heuristic consideration is the appropriate threshold for a robustness claim. How different do the heuristics need to be in order to be justified in claiming that a shared result is a robust result? Levins' answer was an appeal to independence:

> Γ_2: To achieve robust results, sample from models or experiments that use independent heuristics.

This is a more specific claim than Γ_1, since it does not just prescribe that we perturb our heuristics, we must now ensure that our heuristics are independent. But what sort of "independence" could this be? Orzack and Sober (1993) are skeptical that there can be any meaningful independence requirement. Finding that neither logical nor statistical concepts of independence fit the bill, they write:

> Robust theorems based on "independent" models would be desirable if we could get them. . . . We have no magic test procedure that investigators can use here, only the caveat that the value of robustness depends on the models' degree of independence. This latter quantity, unfortunately, is elusive (Orzack and Sober 1993, p.540).

They go on to briefly consider robustness without an independence requirement, but this leads them only to the unhappy situation in which "[e]very statement about nature is robust in this

[4] I will have more to say on the relationship between our understanding of a mathematical model and our understanding of the natural world in the next chapter.

sense because every statement is entailed by more than one model" (Ibid.).[5]

Setting aside Orzack and Sober's attempt to trivialize the concept, the deeper worry about being able to distinguish genuine properties of robustness from pseudo-robustness is important. Levins and Wimsatt both write of "model independence" or "independent tests" to achieve robustness, but apart from a weak independence requirement that a model or test should neither imply nor rule out its result a priori, it is hard to make sense of what independence would be. Even Weisberg, who champions the values of robustness, explicitly admits that there will be a common theoretical core or structure to the sampled models of a robustness analysis (Weisberg 2006, p.737).

Strong independence of models may thus be unobtainable. But we do not need to insist on so much. Instead, I argue that what matters for robustness is a *diversity* of heuristics. As I suggested of Levins' example, the three sampled models made use of two different heuristics, each looking at a single level of organization. These two heuristics are not only independent (in the weak sense), they are actually inconsistent. Nevertheless, if we think that the environment's effect on species polymorphism might importantly take place at multiple levels, then we might judge this to be an inadequately diverse set of heuristics.

However, substituting the term "diversity" for "independence" is not much of an improvement unless we can say something more about what diversity means or how it can be measured in different contexts. Sensitive to this concern, Weisberg writes:

> In straightforward cases, the common structure [among the sampled models] is literally the same mathematical structure in each model. In such cases one can isolate the common structure and, using mathematical analysis, verify the fact that the common structure gives rise to the robust property. However, such a procedure is not always possible because models may be developed in different mathematical frameworks or may represent a similar causal structure in different ways or at different levels of abstraction. Such cases are much harder to describe in general, relying as they do on the theorist's ability to judge relevantly similar structures. In the most rigorous cases, theorists can demonstrate that each token of the common structure gives rise to the robust behavior and that the tokens of the common structure contain important mathematical similarities, not just intuitive qualitative similarities. However, there are occasions in which theorists rely on judgment and experience, not mathematics or simulation, to make such determinations (Ibid., p.738).

[5] The complaint that investigators "have no magic test procedure" to judge robustness is a common one, and seems to reflect a preoccupation with making scientific judgments as much like mathematical operations as possible. As I showed in the last chapter, a similarly spirited attempt to "solve" Duhem's problem with inductive logic, thereby reducing scientific judgments to an epistemic calculus, was as unilluminating as it was unnecessary. I see no reason why the techniques to analyze robustness cannot be as varied and numerous as the models and heuristics themselves.

My analysis in the last section rules out what Weisberg calls "straightforward cases" as relevant only to stability analyses, rather than robustness. Nevertheless, his concluding remarks referring to "occasions in which theorists rely on judgment" speaks to the task at hand. On my view, this is a tacit appeal to meta-heuristics. Determining what counts as a sufficient exploration of the problem space—sufficient to ensure that we have found a reliable result—requires that we have at least some meta-heuristics describing the relevant dimensions and explored regions. Because a problem space may be conceptualized in different ways and at different levels of decision-making, figuring out how best to describe it is itself a meta-heuristic consideration. But bearing this complexity in mind, I will now discuss one way of going about it, using an example from the science of clinical trials.

2.2.1 Diversity of pragmatic and explanatory trial designs

There is not simply one way to go about designing a clinical trial. Clinical research is a complex web of experiments, requiring investigations into basic biology, pharmacology, animal diseases—and all this before even getting into research with human subjects. Once human trials begin, the research questions further multiply, requiring that we ask about safety, tolerability, human pharmacokinetics, treatment efficacy, and then finally, treatment effectiveness in clinical practice.

Toward the end of this long chain of research, in the so-called "phase II" and "phase III" human trials, the pressing questions deal with a treatment's efficacy—its specific biological effect on the condition of interest—and its effectiveness—its performance in conditions similar to clinical practice. At this stage, a distinction is commonly drawn between two kinds of trials: explanatory trials and pragmatic trials. Explanatory trials are highly controlled experiments and well-suited to answering questions about efficacy. Pragmatic trials are less controlled experiments and seek to simulate conditions more like those of clinical practice and well-suited to answering questions about effectiveness.[6]

Thorpe et al. (2009) observe that the explanatory/pragmatic distinction is not absolute. Most trials are a mixture of explanatory and pragmatic assumptions. To represent the spectrum of trial assumptions, they propose a methodological tool called PRECIS: the PRagmatic-Explanatory Continuum Indicator Summary, writing:

> Very few trials are purely pragmatic or explanatory. For example, in an otherwise explanatory trial, there may be some aspect of the intervention that is beyond the investigator's control. Similarly, the act of conducting an otherwise pragmatic trial may impose some

[6] See chapter 4 for a critical examination of these concepts and their implications for the epistemology and ethics of clinical research.

> control resulting in the setting being not quite usual. For example, the very act of collecting data required for a trial that would not otherwise be collected in usual practice could be a sufficient trigger to modify participant behavior in unanticipated ways. Furthermore, several aspects of a trial are relevant, relating to choices of trial participants, health care practitioners, interventions, adherence to protocol, and analysis. Thus, we are left with a multidimensional continuum, rather than a dichotomy and a particular trial may display varying levels of pragmatism across these dimensions (Thorpe et al. 2009, p.465).

Thorpe et al. propose ten dimensions along which trials may slide either toward an explanatory or pragmatic design strategy. These dimensions, along with extreme examples of the respective pragmatic or explanatory assumptions, are summarized in table 2.1. As the table makes explicit, the designing trialists are faced with an array of choices, each of which has an impact on the inferences that can later be justified by the trial's result. A trial structured toward the explanatory end of the spectrum, utilizing more eligibility restrictions, stricter protocol, etc., may have a high degree of internal validity, meaning that we can be very confident about its result as it applies to the study population. However, we should be less confident about its external validity; cautious about inferring that its result will apply to the patient population at large. Inasmuch as the trial was unlike clinical practice, there may be reasons to doubt that the treatment will be effective there.

A trial structured toward the pragmatic end of the spectrum has just the opposite problem. Since it will more closely resemble clinical practice, the data collected will tend to be "noisier", and it will be more difficult to determine if the outcomes measured can truly be attributed to the biological effects of the treatment. In this sense, its internal validity is weaker, but its external validity—its generalizability to the larger population—will be stronger.

In light of this trade-off on validity and justified inferences, we might think that to establish a robust result in clinical research, both explanatory and pragmatic trials ought to play a role.[7] We want to be certain of the biological effects of a new treatment, and for this, trials to the explanatory end of the spectrum will be most appropriate. Once we are confident that a treatment has a genuine effect, we want to be sure that it will work in clinical practice to treat the general population. For this, pragmatic trials are needed. And within those studies, it might also be wise to perturb some of the different dimensions, adding for example, a more explanatory trial

[7] The trade-off between explanatory trial design and internal validity, on the one hand, and pragmatic trial design and external validity, on the other, reflects the standard reading of internal validity: Controlling for more possible confounders eliminates other potential explanations of the result, and thereby increases the strength of the causal inference. However, an alternative understanding of internal validity, as meaning a match between the study question and the study methods, presents no such trade-off for a more pragmatic trial. A pragmatic trial may simply investigate a different hypothesis; content to demonstrate a relationship between treatment assignment and improvement. Therefore, it need not rule out as many confounding factors in order to possess a high degree of internal validity.

Table 2.1: PRECIS dimensions

Dimension	Pragmatic Trial	Explanatory Trial
Participants eligibility criteria	All participants who have the condition of interest are enrolled, regardless of their anticipated risk, responsiveness, co-morbidities, or past compliance.	Stepwise selection criteria are applied that: (a) restrict study individuals to just those previously shown to be at highest risk of unfavorable outcomes, (b) further restrict these highrisk individuals to just those who are thought likely to be highly responsive to the experimental intervention, and (c) include just those high-risk, highly responsive study individuals who demonstrate high compliance with pretrial appointment-keeping and a mock intervention.
Experimental intervention flexibility	Instructions on how to apply the experimental intervention are highly flexible, offering practitioners considerable leeway in deciding how to formulate and apply it.	Inflexible experimental intervention, with strict instructions for every element.
Experimental intervention practitioner expertise	The experimental intervention typically is applied by the full range of practitioners and in the full range of clinical settings, regardless of their expertise, with only ordinary attention to dose setting and side effects.	The experimental intervention is applied only by seasoned practitioners previously documented to have applied that intervention with high rates of success and low rates of complications, and in practice settings where the care delivery system and providers are highly experienced in managing the types of patients enrolled in the trial. The intervention often is closely monitored so that its "dose" can be optimized and its side effects treated, and co-interventions against other disorders often are applied.
Comparison intervention	"Usual practice" or the best available alternative management strategy, offering practitioners considerable leeway in deciding how to apply it.	Restricted flexibility of the comparison intervention and may use a placebo rather than the best alternative management strategy as the comparator.
Comparison intervention practitioner expertise	The comparison intervention typically is applied by the full range of practitioners, and in the full range of clinical interest, regardless of their expertise, with only ordinary attention to their training, experience, and performance.	Practitioner expertise in applying the comparison intervention(s) is standardized so as to maximize the chances of detecting whatever comparative benefits the experimental intervention might have.
Follow-up intensity	No formal follow-up visits of study individuals at all. Instead, administrative databases (such as mortality registries) are searched for the detection of outcomes.	Study individuals are followed with many more frequent visits and more extensive data collection than would occur in routine practice, regardless of whether they had suffered any events.
Primary trial outcome	The primary outcome is an objectively measured, clinically meaningful outcome to the study participants. The outcome does not rely on central adjudication and is one that can be assessed under usual conditions: for example, special tests or training are not required.	The outcome is known to be a direct and immediate consequence of the intervention. The outcome is often clinically meaningful, but may sometimes (early dosefinding trials, for example) be a surrogate marker of another downstream outcome of interest. It may also require specialized training or testing not normally used to determine outcome status or central adjudication.
Participant compliance with "prescribed" intervention	There is unobtrusive (or no) measurement of compliance, and no special strategies to maintain or improve compliance are used.	Study participants' compliance with the intervention is monitored closely, may be a pre-requisite for study entry, and both prophylactic strategies (to maintain) and "rescue" strategies (to regain) high compliance are used.
Practitioner adherence to study protocol	There is unobtrusive (or no) measurement of practitioner adherence and no special strategies to maintain or improve it are used.	There is close monitoring of how well the participating clinicians and centers are adhering to even the minute details in the trial protocol and "manual of procedures".
Analysis of primary outcome	The analysis includes all patients regardless of compliance, eligibility, and others (the "intention-to-treat" analysis). In other words, the analysis attempts to see if the treatment works under the usual conditions, with all the noise inherent therein.	An intention-to-treat analysis is usually performed; however, this may be supplemented by a per-protocol analysis or an analysis restricted to "compliers" or other subgroups to estimate maximum achievable treatment effect. Analyses are conducted that attempt to answer the narrowest, "mechanistic" question (whether biological, educational, or organizational).

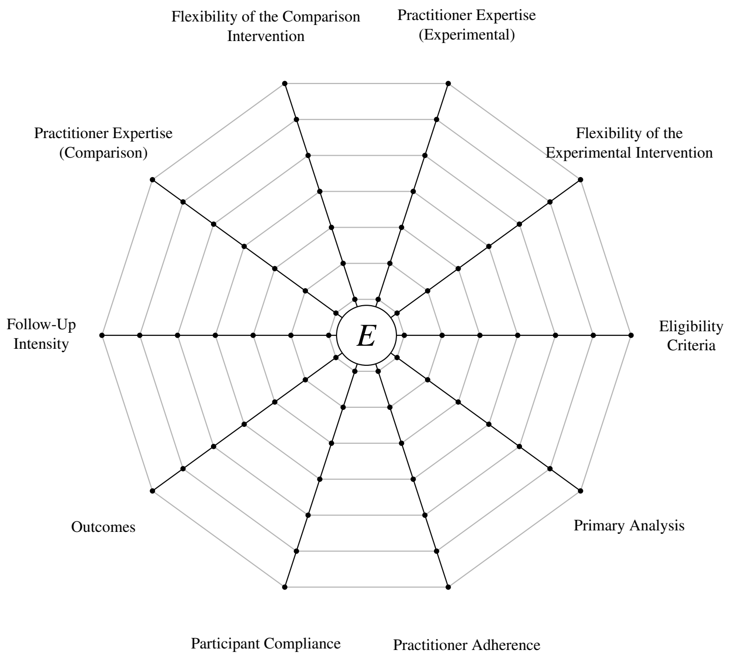
Flexibility of the Comparison Intervention
Practitioner Expertise (Experimental)
Practitioner Expertise (Comparison)
Flexibility of the Experimental Intervention
Follow-Up Intensity
E
Eligibility Criteria
Outcomes
Primary Analysis
Participant Compliance
Practitioner Adherence

Figure 2.2: Blank PRECIS "wheel"

that does not restrict its study population with excessive eligibility criteria or a more pragmatic trial that includes intense follow-up.

These robustness considerations can be represented visually using Thorpe et al.'s PRECIS "wheel", a specialized radar graph, shown in figure 2.2. The graph is a representation of the relevant problem space of trial design, specifying the various dimensions of perturbation. The extent to which a trial's assumptions along these dimensions is either explanatory (toward the *E* in the center of the wheel's axis) or pragmatic (towards its outer edge) can be plotted along the chart, and each plot thereby represents a particular exploration of the problem space of clinical trial design.

But the PRECIS graph offers more than mere representation, as Thorpe et al. write:

> The simple graphical summary [provided by the PRECIS wheel] is a particularly appealing feature of this tool. We believe it to have value for the planning of trials and assessing whether the design of a trial is fit for purpose. It can help ensure the right balance is struck to achieve the primary purpose of a trial, which may be to answer an "explanatory" question about whether an intervention can work under ideal conditions or to answer a "pragmatic" question about whether an intervention does work under usual conditions. PRECIS highlights the multidimensional nature of the pragmatic-explanatory continuum. This multidimensional structure should be borne in mind by trial designers and end-users alike so that overly simplistic labeling of trials can be avoided (Thorpe et al. 2009, p.474).

In addition to providing an important check between the research question and design of a sin-

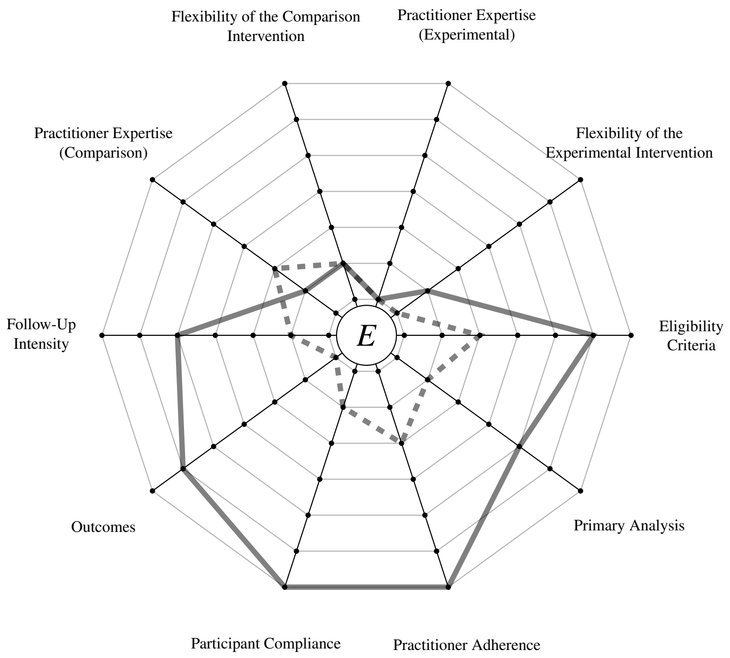

Figure 2.3: Example PRECIS graph for 2 trials

gle study, one can also use the graph to analyze how thoroughly a series of trials has explored the problem space, i.e., provide a visual representation of diversity amongst the trial design heuristics. Figure 2.3, for example, illustrates how one could "overlay" the plots of two different trials onto the same graph. In the figure, one of these trials is very explanatory (dashed line), the other is a mixture of pragmatic and explanatory assumptions (solid line). Supposing that both trials found a positive result with the experimental treatment, we have some evidence in favor of a robust result. Nevertheless, a large area of the graph remains unexplored, particularly in the regions for flexibility of application and practitioner expertise. Just as a single plot provides an immediate visual impression of the trial's overall strategy, a layered plot can provide an immediate impression of how well-explored is the space of possible trials.

But we must be careful about what this representation shows. Although the area of the graph enclosed by the more pragmatic trial's plot is significantly larger, we should not mistake this as meaning that a pragmatic trial can thereby "cover" more of the possible strategies. Indeed, the choice of representing the more explanatory end of the design spectrum at the center of the graph is arbitrary. It could just as well have been the edge. Therefore, the enclosed area of the a plot does not represent a legitimate epistemic property.

On the other hand, the "nearness" of two plots on the graph can reflect a legitimate experimental relationship. The solid plot and the dashed plot overlap along the dimensions of practitioner expertise (experimental) and flexibility of comparison intervention, and are only one step away along the dimensions of practitioner expertise (comparison) and flexibility of experimental intervention. This should indicate that the two trials were indeed identically or similarly explanatory in their assumptions along this dimension. Whereas, the much greater "distance" between the two trials along the dimensions of outcomes and participant compliance indicates a significant diversity of design.

Each plot explores only the area directly underneath its path, but it can be taken as a sample for the nearby regions, which are assumed to be similar. Thus, it would be largely uninformative to run a trial which is only slightly more explanatory or less pragmatic than another along one or a few dimensions. Conversely, the more a trial perturbs its assumptions along these dimensions, relative to previous studies, the more it is potentially informative.

In an extreme and unrealistic example, we can imagine two studies of unimpeachable quality: one is an exclusively explanatory trial, the other exclusive pragmatic. Both show strong positive results for a treatment's efficacy and effectiveness, respectively. This could provide some justification for thinking that the problem space was largely uniform, and had therefore been adequately explored after only two studies. In other words, the treatment was shown to be both efficacious and effective, so we need not continue with further trials.

The much more common case will, of course, be variable performance of a treatment across studies of mixed pragmatic-explanatory dimensions. This makes judgments about robustness, dependent upon an understanding of the overall problem space, more difficult. Each individual study provides only a limited amount of information, and may be subject to questions about the quality of its evidence. Balancing the demands of robustness with the realities of clinical research is a complex issue; one that I will return to in chapters 4 and 5.

The negative consequence of the large, unexplored region in figure 2.2, is that judgments about the robustness of the intervention's effectiveness—how it would perform in practice—should be tentative. The lack of flexibility of the experimental and comparator interventions and the demand for high practitioner expertise in both of these trials leaves us without any evidence for how the intervention might perform when these conditions are relaxed. The positive consequence is that it suggests how a future study might be designed. Since a robust demonstration of the treatment's effectiveness is the ultimate aim, then this PRECIS graph of the problem space reveals that there is not yet a *sufficiently diverse set* of experiments. Thus, we can appeal to the meta-heuristic strategy of robustness, Γ_1, recognizing the need to perturb future designs in order to explore this region.

For the purpose here, it is most important to see that PRECIS's representation of problem space—a scaled radar graph—need not be limited to the analysis of clinical trials. The dimensions, scales, and derivative properties of the representation (e.g., what a plot near center of the graph implies versus a plot nearer the outer edge) will obviously be science or investigation specific. Nevertheless, the uses of such a radar graph are potentially applicable and illuminating for understanding robustness considerations more broadly; sharpening meta-heuristic claims about relevant parameters and sufficient heuristic diversity.

Constructing an informed and practically useful representation of problem space also addresses each of the four activities of Duhemian good sense (cf. S 1.1): The ten dimensions selected by Thorpe et al. are basic conceptual assumptions about the relevant factors in clinical trial design (i.e., activity 1). Once chosen, these dimensions specify an overarching structure of trial design by which specific design judgments will be made, analyzed, and interpreted (i.e., activities 2 and 3). Finally, Thorpe et al.'s awareness of PRECIS as a work-in-progress fits nicely into the fourth activity of good sense—working to achieve scientific consensus (i.e., activity 4):

> Because trials are typically designed by a team of researchers, PRECIS should be used by all involved in the design of the trial, leading to a consensus view on where the trial is situated within the pragmatic-explanatory continuum. The possible subjectiveness of dot placement should help focus the researcher's attention on those domains that are not as pragmatic or explanatory as they would like. Clearly, those domains where consensus is difficult to achieve warrant more attention (Ibid., p.474).

2.2.2 Diversity of biological modeling

So how might the PRECIS tool be modified and extended for other sciences? Recall that on Levins' account, modelers must make trade-offs between generality, realism, precision, manageability and understandability. Let us take these as the dimensions of a new problem space for biological modeling. Although he does not define each of these properties and their consequences as thoroughly as we saw Thorpe et al. do for trial design decisions, we can try to take this step for him. Table 2.2 presents the five modeling dimensions with examples of the "maximum" or "minimum" extremes.

Using these dimensions, Levins descibes three possible strategies; meta-heuristics of model building:

Γ_3: Sacrifice generality for the sake of realism and precision.

Γ_4: Sacrifice realism for the sake of generality and precision.

Γ_5: Sacrifice precision for the sake of generality and realism.

Taking these three strategies, we can then construct and overlay plots in a problem space of biological modeling, like that shown in figure 2.4. In the figure, Γ_3 is represented by the solid line; Γ_4 by the dotted line; and Γ_5 by the dashed line. Although it is implicit in Levins'

Dimension	Maximum	Minimum
Generality	Models are a simple as possible, including only those parameters thought to represent universal (and usually fundamental) behavior of the entities in question.	Models include all and only those parameters that are relevant to the characteristic, short-term behavior of the specific organism or system of study.
Realism	Models will be carefully derived from field work, including as many known, relevant parameters as understandability will permit.	Models include a number of highly idealized assumptions in order to provide a simplified analysis, with the hope that as a natural system deviates from the model, it will suggest areas in which to complicate the model further.
Precision	Models will be concerned to capture the immediate, short-term behavior of the system. Predictions will be discrete and deterministic.	Models will be concerned with the long-term, qualitative behavior of the system. Predictions will typically be in the form of inequalities.
Manageability	Models will be mathematical or computer simulations. Initial conditions and parameters can be varied at will.	Models will be derived from observational or historical studies where the modelers must infer the specific initial configurations and parameters that may have caused the behavior of the system in question.
Understandability	Models will be able to establish stability properties, demonstrating clear relationships among the different variables and parameters.	Models will be attempts to deal with irreducibly complex case-studies. Explanations will only ever be partial, with any results remaining susceptible to multiple, possibly contradictory, interpretations.

Table 2.2: Biological modeling dimensions

characterization, the resulting picture explicitly shows how each of these strategies is only a partial exploration of the space. But just what implications that has for modeling strategies and biological explanations is not yet clear.

Thorpe et al.'s problem space is elucidated by the difference in research questions and inferences between pragmatic and explanatory trials, which establishes the epistemic connection between the PRECIS representation and the aims of the trial. It is also at what we might think of as a lower methodological level, representing specific assumptions in the design of a clinical trial. This is akin to the choice of decision-points in the fast-and-frugal model discussed in §1.3.2; representing the *content* of judgments, rather than the heuristics.

Levins' dimensions of modeling are more abstract, operating at the heuristic level. His different strategies, $\Gamma_3 \ldots \Gamma_5$, prescribe that we sacrifice a particular property of the model, but it does not tell us the content of this property for any given problem. As a result, it is less clear how he thinks each strategy ought to be used and what inferences it can justify. He does write that he prefers the third strategy (i.e., sacrifice precision) and suggests that he looks upon the second (i.e., sacrifice realism) with some disdain; nevertheless a more complete articulation of the problem space for strategies in biological modeling would require additional meta-heuristics to make relationships like that between strategy plots and their well-justified inferences explicit.

Additional argument may also be needed to justify Levins' choice of dimensions. Just as Thorpe et al.'s decision to use the particular ten aspects of trial design was a collaborative process and an on-going negotiation with the scientific community, so too should Levins' dimensions be considered targets for revisions. What and how many dimensions need to be

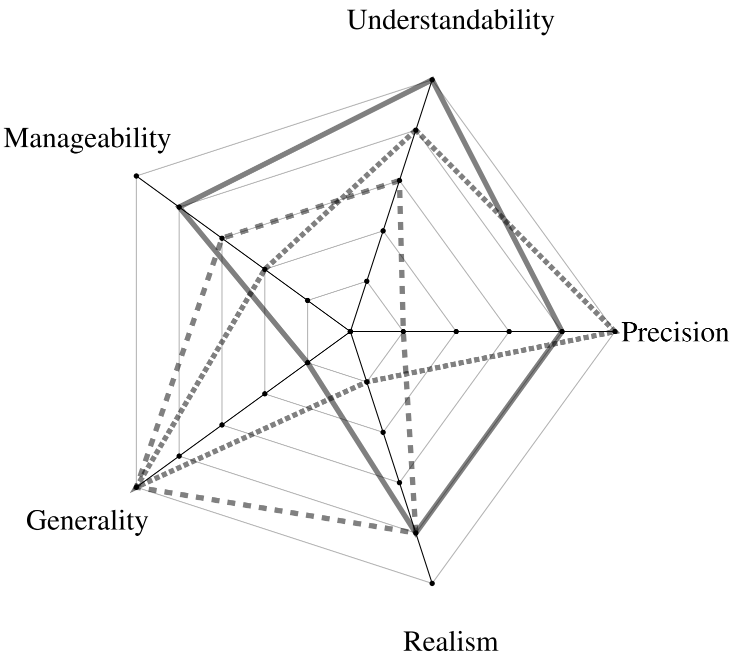

Figure 2.4: Problem space for modeling strategies in population biology

represented is a meta-heuristic consideration. Just as modelers must decide how to appropriately simplify their models to address a question, so must the scientist (or scientific community) decide how to characterize the problem space to adequately inform judgments about justified inferences or robustness.

This is another way in which a representation of problem space must work in harmony with the judgments of good sense. Dimensions which initially seem important or essential may come into question. For example, Levins' dimension of manageability may no longer be as relevant as computational power increases. Or new dimensions may need to be added as the domain of the question grows larger. For example, Levins' modeling dimensions do not obviously incorporate features of laboratory or field models, and these features may be essential to judging that a biological result is robust.

Refining our understanding of the problem space thus becomes a central epistemic and methodological project as research is ongoing. With respect to robustness, and the narrower aims of this chapter, it suffices to say that a radar graph representation of problem space, elucidated by meta-heuristics for the choice of dimensions, scales, and derivative properties, can show how multiple means of investigation provide the needed diversity. A result shared by multiple models, which combine to adequately explore the relevant problem space, demonstrates the robustness property. Moreover, the aim of achieving this property—leveraging an understanding of the space to see how the heuristics and assumptions need to be perturbed in order to more fully explore the space—elucidates the meta-heuristic sense of robustness as a strategy.

2.3 Discordance and Relevance

The philosophical literature has some discussion of counter-examples to the use or value of robustness. Orzack and Sober (1993), for example, appeal to Wade's (1978) review of group selection models in population biology as one such counter-example. In that case, twelve mathematical models were taken to show that group selection would not be efficacious in natural contexts— a seemingly very robust result. Yet, Wade argues that these models all shared a common set of assumptions, which biased them against group selection's efficacy. Orzack and Sober take this to be evidence that robustness cannot be principally distinguished from pseudo-robustness. Wimsatt (1982, 2007) takes a different view of that case, arguing that Wade's review is actually evidence of *how* robustness and pseudo-robustness are differentiated.[8]

Rasmussen (1993) offers another supposed counter-example with the case of the illusory mesosomes. The mesosome was a bag-shaped membranous structure "observed" by cell-biologists using an electron microscope and a variety of preparation techniques. After fifteen years of work on mesosomes, it was finally discovered that they were mere artifacts; the result of a bias shared by all of the preparation techniques. Like Orzack and Sober, Rasmussen takes this to be evidence that robustness cannot be distinguished from pseudo-robustness. He is met by Culp (1994), who, like Wimsatt, insists that robustness is ultimately the reason why the mesosome was finally determined to be an artifact. She argues that the evidence for its artifactuality is *more* robust than the evidence for its reality.

Stegenga (2009) is aware of both of these cases, but offers an entirely different line of critique. He raises two conceptual objections against any principled means of determining robustness. As I noted at the outset of the chapter, he defines robustness as: "[t]he state in which a hypothesis is supported by evidence from multiple techniques with independent background assumptions" (Stegenga 2009, p.651), and is thus focused on the property of robustness and whether we can legitimately judge it.

The first problem he identifies is that of discordant data, which he breaks down into the two sub-categories of inconsistency and incongruence:

> Inconsistency is straightforward: petri dishes suggest x and test tubes suggest $\neg x$. In the absence of a methodological metastandard, there is no obvious way to reconcile various kinds of inconsistent data. Incongruity is even more troublesome. How is it even possible

[8] I share Wimsatt's view of this case, and discuss it in some depth in chapter 3.

> for evidence from different types of experiments to cohere? Evidence from different types of experiments is often written in different 'languages'. Petri dishes suggest x, test tubes suggest y, mice suggest z, monkeys suggest $0.8z$, mathematical models suggest $2z$, clinical experience suggests that sometimes y occurs, and human case control studies suggest y while randomized control trials suggest $\neg y$ (Stegenga 2009, p.654).

He observes that each "translation" of a result between these "languages" requires background assumptions, linking the conceptual domain of one field to that of another. Such background assumptions may have varying degrees of plausibility, and if any link is implausible, then the contribution from that mode of evidence toward robustness is undermined. "[R]obustness-style arguments presuppose a principled and systematic method of assessing and amalgamating multimodal evidence, and without such methods of combining evidence, robustness arguments are merely intuitive or qualitative" (Ibid., p.655).

Reminiscent of Orzack and Sober's criticism, that there is no "magic test procedure" to show us that we have a robust result, Stegenga's worry that robustness may turn out to be "merely intuitive or qualitative" is, I argue, similarly misplaced. He assumes a distinction between the "principled and systematic" methods, which we are to understand as rigorous and quantitative, and the "merely intuitive or qualitative" methods. This is simply a false dichotomy. There can be principled, systematic, *and yet* qualitative methods. Thorpe et al.'s approach to the pragmatic-explanatory continuum is one such. There can also be mixed quantitative and qualitative methods—as Levins takes his third modeling strategy to be.

I am happy to admit that robustness arguments must be qualitative. However, this is no reason to suppose that they are any weaker or less useful because of it. As I argued in §2.1, the quantitative analyses of stability theory (which is often called "robustness") has certain advantages due to its rigor and precision, but it also has certain disadvantages. A stability analysis cannot, without further assumptions and argument, justify claims to reliability in the natural world. For the pure mathematician, this is obviously not an issue. But for the scientific modeler, who wishes to extend his result from the space of mathematical possibilities to the space of natural systems, this additional step is critical. And in every science, except perhaps theoretical physics, it will be a qualitative step.

Stegenga's claim that robustness arguments must be "intuitive" is also consistent with the overarching project of this thesis. Explicating scientific judgment and the meta-heuristics of good sense is the work of spelling out such "intuitive" methods. Robustness is a central methodological component of this picture and the aim of this entire chapter has been to clarify its "intuitive" use and value. Therefore, the charge of "intuitive" or "qualitative" is not damaging.

Finally, Stegenga's observation that "multimodal evidence, when available, is rarely concordant" (Ibid.) overstates the case significantly. One of Wimsatt's basic examples of multimodal robustness is the use of different sensory modalities to detect the same property (Wimsatt 2007, p.45). Indeed, we might think of this as the most basic kind of robustness. I see a table in front of me and it looks solid. I touch the table and it feels solid. Obviously, multimodal evidence from scientific experiments are more sophisticated than that homely example. But it is a kind of concordant, multimodal evidence, and even if it is only a part of the experimental background assumptions, it is anything but rare in science.

We might interpret Stegenga more charitably as claiming that there is no shortage of recalcitrant scientific results and that different kinds of results will often be difficult to amalgamate. But this amounts to nothing more than a corollary of underdetermination. Yes, our scientific judgments about how to handle evidence are underdetermined. This is where good sense comes into play. We have heuristics and meta-heuristics to help us sort the evidence out. And for those times when we yet lack heuristics and meta-heuristics to make a well-informed and principled judgment, we adapt to use what we have. If we fail, we re-evaluate our decision tools and refine the problem space.

This is where we can appeal to the *strategy* of robustness. There are no guarantees that we will find the *property* of robustness, and Stegenga's worries about discordance reinforce that point. But this is not a problem for the strategy of robustness. Robust results may prove elusive, but this does not mean that we should stop looking for them.

Stegenga's second objection is the problem of relevance:

> [S]ome evidence must be selected as relevant from discordant data generated by multiple kinds of techniques. Which kinds of data are most relevant to the hypothesis? Which kinds of data are high-quality? Scientists and policy makers can consider data from all kinds of experiments, or only data from high-quality experiments, or only data from one particular kind of experiment. How should they choose? (Ibid., p.658)

He acknowledges that while there is no universal standard in science, there are some science-specific standards for deeming certain kinds of evidence to be of high quality, and therefore, more relevant to considerations of robustness. For example, in the evidence-based medicine movement, the randomized controlled trial is considered to be the gold-standard of evidence, followed by prospective cohort studies, case control studies, observational studies, and case reports (Ibid., p.659). A principled way of judging relevance could also help to resolve the problem of discordance, but again, he thinks that such techniques are, at best, disputed.

As with the problem of discordance, Stegenga's worry here is due to his exclusive focus on robustness as a property of hypotheses. It is true that the property of robustness is not rigorously measurable. As a result, there is no universal, mechanical way to determine that some techniques are more reliable than others, which would in turn allow us to claim that evidence from those more reliable techniques was of higher quality, greater relevance, etc. But once we step away from the property of robustness to think also about robustness as one of the meta-heuristics of research, this problem is not in the least fatal.

Determining what evidence is relevant to a problem is a basic component in constructing a problem space. As I argued in the last section, getting the right dimensions is a meta-heuristic consideration, and it will be investigation specific. Again, we should not expect there to be universal standards applicable to all sciences and kinds of evidence. Instead, we should expect to find heuristics and meta-heuristics in specific disciplines; time-tested *procedures* for separating the high-quality evidence from the poor-quality evidence. Thorpe et al.'s PRECIS tool performs this function for a certain class of methodological questions in clinical research. A problem space for Levins strategies of model building, once supplemented with further meta-heuristics, could perform a similar function for population models in biology.

Stegenga acknowledges that "the prescription to get more data, from different kinds of experiments, is not something that scientists need to be told; they already follow this commonsense maxim" (Ibid., p.655). It is worth considering why this maxim (related to the meta-heuristic Γ_1) is so widespread. It is not because scientists are naïve empiricists, ignorant of the challenges posed by discordant or irrelevant data. It is because the search for robustness and the technical improvements brought about through that search are invaluable. Discordance and relevance are two obstacles that stand in the way of achieving robustness, but they do not thereby undermine its value.

Chapter Summary

The aim of this chapter has been to establish the prominent role that robustness, both as property and as strategy, must play in the framework of good sense. Once its role is distinguished from that of mathematical stability (along the lines of figure 2.1), and we come to recognize its relationship to heuristics, we are then well-equipped to make sense of the practice and judgments involved with multi-modal techniques and evidence: A robust result, reliable for understanding natural phenomena, is one that is invariant across models utilizing a diversity of heuristics. Therefore, we ought to adopt the meta-heuristic, Γ_1, which prescribes that we strive to perturb heuristics across the space of possible investigations.[9]

[9] There is a qualification to make here between scientific domains that must use partial models and domains that are characterized by a single model. The domain of population biology certainly falls into the former category, since, as Levins observed, a complete model would so impossibly complex as to be entirely unilluminating. The domain of gravitation, on the other hand, could arguably fit into the latter category, relying solely on the inverse square model. At first blush, it might seem as though my account of robustness would fail to apply in

But there can be no guarantees of robustness. Concordant results may prove elusive or we may be deceived by hidden biases or inadequate diversity of heuristics. Thus, we must also work to develop epistemic tools—heuristic means for evaluating robustness claims. The PRECIS tool developed in clinical trials, overlaying the plots of multiple trials onto a radar graph of experimental dimensions, offers one way of representing and analyzing diversity in a specific scientific context. I have suggested how this tool might also be extended for use in biological modeling. A more thorough discussion and treatment of these two problem spaces will be provided in the subsequent chapters.

Tools like the PRECIS graph help to make the meta-heuristic dimensions of a problem explicit. They reflect judgments about the important features of a model or experiment and can aid in the interpretation of results. Do our methods match our research aims? Has the problem space been adequately explored? Have we perturbed our models and experiments in an optimal way? These are the kinds of important meta-heuristic questions lying behind robustness claims. I have argued that they can all be illuminated by the framework of good sense.

such single-model domains, given that the heuristics of model construction will be constant. But heuristics of model construction are not the only kinds of heuristics. Heuristics of conceptualization, experimentation, and interpretation may still be perturbed in evaluating the predictions of even a single model. While a thorough exploration of this distinction would take us too far afield, it is well worth acknowledging that there may be degrees to which a scientific domain admits of robustness analysis.

Chapter 3

Meta-heuristics in the Group Selection Controversy

Throughout much of the 20th century, biologists have debated whether groups of organism can be units of natural selection. Lloyd's (2005) response to the reductive arguments in Sterelny and Kitcher (1988) and Kitcher, Sterelny, and Waters (1990) suggests that although the terms of the controversy have changed, the core questions remain open: What kinds of things does natural selection act on? And at what levels of organization ought we to invoke natural selection as an explanation?

In this chapter, I return to what might be called the "first wave" of the group selection controversy—the late 1960s through the early 1980s. Williams' *Adaptation and Natural Selection* (1966) is often thought to have struck a "fatal blow" against the proponents of group selection in biology at that time. As D. S. Wilson recalls it, this effect came "not from a crucial experiment, or even from a new theoretical development, but simply from the elegance and clarity of Williams' thought in interpreting developments of the previous three decades" (Wilson 1983, p.159).

Part of Williams' "clarity of thought" can be understood as articulating a heuristic of biological explanation. He writes:

> The ground rule—or perhaps the *doctrine* would be a better term—is that adaptation is a special and onerous concept that should be used only where it is really necessary. When it must be recognized, it should be attributed to no higher a level of organization than is demanded by the evidence. In explaining adaptation, one should assume the adequacy of the simplest form of natural selection, that of alternative alleles in Mendelian populations, unless the evidence clearly shows that this theory does not suffice . . . (Williams 1966, p.4, original emphasis)

Let us summarize Williams' view with the following heuristic:

> β_1: Selection should be attributed to no higher a level of organization than is demanded by the evidence.

He argues that once a clear distinction is made between a population of *adapted organisms*, on the one hand, and an *adapted population* of organisms, on the other, we will see that the appearance of the latter is typically nothing more than the statistical summation of the former.

Williams fills his analysis with examples of purported group selection than can also be explained with individual selection. For example, he considers the feeding activities of earthworms. It so happens that the earthworm's digestion improves the quality of the soil, which benefits both the entire population of earthworms and the surrounding ecological community. However, he cautions that it would be an error to think of the earthworm's digestive system as an adaptation for soil improvement. If we accept β_1, then we ought to explain the earthworm's digestive system as an adaptation for individual nutrition. No explanatory appeals to group selection are demanded; and therefore, none should be made.

The economy of β_1 is certainly indisputable, and the distinction between adapted organisms and adapted populations is a valuable insight. The problem with β_1 is that it is obviously dependent upon empirical considerations: Are there, in fact, observed adaptations in nature that require an explanation by group selection? What is the force of "demanded by the evidence"? To put it in the language of this thesis: What is β_1's appropriate region of problem space?

It will be instructive to examine the surrounding research context of Williams' heuristic, to see if it *should* have held up as it did in the face of experiment and theoretical development. To that end, this chapter will focus on three arguments in the group selection debate, each of which enriches our understanding of the overall problem space. I begin, in the next section, with the mathematical group selection models of Wright (1945) and Maynard Smith (1964). Both of these models admitted the possibility of efficacious (and therefore explanatory) group selection, but were taken to show that the parameter values needed to produce group selection were too extreme to be found in nature. I argue that this conclusion was unjustified. Building on the work of the last chapter, I show how this conclusion mistakes the inferential properties of mathematical stability for those of robustness.

In §3.2, I discuss Wade's (1976, 1977) experimental work. He is able to demonstrate group selection's efficacy in the laboratory, a result that potentially restricts the domain of β_1 and challenges the relevance of the mathematical structures assumed in the models of Wright, Maynard Smith, and others (cf. Levins 1970; Eshel 1972; Boorman and Levitt 1973; Levin and Kilmer 1974; Wilson 1975, 1977; Gadgil 1975; Charnov and Krebs 1975; Gilpin 1975; Matessi and Jayakar 1973, 1976; Cohen and Eshel 1976).

Wade (1978) also reviews the assumptions across twelve group selection models and finds that the models all share a common core of idealizations, which bias them a priori against group selection's efficacy. He proposes alternative assumptions—still idealizations—that are more neutral toward the effects of group selection. I argue that Wade's analysis and proposed modeling idealizations can be considered alongside the twelve earlier mathematical models as contrasting heuristic strategies, in need of some meta-heuristic guidance. Specifically, there is a critical need to distinguish between benign idealizations, which are useful and necessary given a particular problem, from malignant ones, which obscure important features in the object of study.

Finally, in §3.3, I will discuss Griesemer and Wade's (1988) explanatory "pathways". They argue that mathematical models, laboratory models, and field studies justify different inferences in biology: Mathematical models are best used for testing parameter weightings; laboratory models and field studies are useful for explanations in nature. I argue that their prescriptions, along with the schematic "pathways", provide us with a detailed map of the relevant problem space. I will also add to this picture by drawing on Levins' dimensions of biological modeling, discussed at the end of the last chapter. The result is an even richer problem space, useful for classifying investigations, searching for robustness, and judging how future research might proceed.

3.1 Wright's and Maynard Smith's Models

Wright (1945) proposed his model of group selection before Williams introduced β_1, and it is clear that he did not share Williams' dismissive view of group selection. He writes:

> . . . selection between the genetic systems of local populations of a species, operating through differential migration and crossbreeding, has been perhaps the greatest creative factor of all in making possible selection of genetic systems as wholes in place of mere selection according to the net effects of alleles (Wright 1945, p.416).

Wright took his model to explain only one particular kind of group selection process: the preservation of a trait that is advantageous to the population as a whole, but disadvantageous to the individual possessing the trait. For Wright, there were also other kinds of group selection effects in addition to the persistence of "altruism", as it came to be called. For example, he thinks that group selection is needed to explain how a population can transition from one adaptive peak to another (Ibid.). Yet, it is his assumption of individual and group selection operating in opposite directions that became canonical for subsequent models.[1]

[1] As we will see in the next section, this assumption is shared by all twelve of the mathematical group selection modes. Wade's (1976, 1977) experimental work also shows that it is demonstrably false for certain adaptations and population structures.

Genotype	**Frequency**	**Selective Value**
AA	$(1-q)^2$	$(1+bq)$
Aa	$2q(1-q)$	$(1+bq)(1-s)$
aa	q^2	$(1+2bq)(1-2s)$

Table 3.1: Wright's group selection model

Table 1 shows Wright's group selection model, where A is the dominant "selfish" allele, a the "altruist" allele, q the initial frequency of a in the population, b the reproductive benefit bestowed per a allele to each group member by an individual with at least one a allele, and s the cost to the individual per a allele. This gives rise to three phenotypes and genotypes, where Aa and aa are semi-altruists and full-altruists, respectively, bringing reproductive benefit b or $2b$ to all members of their groups, but paying a reproductive cost of s or $2s$ depending upon how many a alleles they carry.

Assuming random mating, the important conclusions are two: First, the presence of the altruist allele in the group increases the group size from one generation to the next so long as:

$$b > \frac{2s}{1-2sq} \tag{3.1}$$

This condition places a constraint on the relevant parameter range for s, since any value of b which does not improve the reproductive success of the group fails to satisfy the primary assumption of the model (i.e., a trait which benefits the group, but is a detriment to the individual).

Second, Δq, the change in frequency of the altruist gene from one generation to the next, is negative:

$$\Delta q = -\frac{sq(1-q)}{1-2sq} \tag{3.2}$$

This means that no matter what the initial frequency of the altruist allele, it will always be declining. It is also important that the benefit parameter, b, does not directly appear in equation 3.2 (although the cost parameter, s, does). D. S. Wilson (1983) notes that:

> According to Wright's model, natural selection is totally insensitive to group benefit, no matter what its value. An analogy might be drawn with a person who buys a number of lottery tickets for himself and the same number of tickets for everyone else in the lottery. That person has not increased his chances of winning the lottery; if he buys more tickets for others than for himself, he will have actually reduced his chances, even though he may have a large number of tickets. Only by obtaining more tickets relative to everyone else can the person increase his chances of winning the lottery (Wilson 1983, p.164).

Despite these consequences, Wright was not ready to give up on group selection. He admits that "the socially favorable mutation [a] tends to be lost or nearly lost in a random breeding population", but invites us to consider a population of many, small, isolated breeding groups, or *demes*, with some amount of migration, m, between them. We can then add the term $m(q_i - q)$ to equation 3.2, where q_i is the frequency of a alleles amongst the migrants. This gives us:

$$\Delta q = m(q_i - q) - (\frac{sq(1-q)}{1-2sq}) \tag{3.3}$$

The background assumption is that larger demes produce more migrants. Since demes with higher q (as a result of drift or sampling error) will grow more quickly, Wright finds it plausible to think that $m(q_i - q)$ "may easily be large enough to overbalance the selective disadvantage of a" (Wright 1945, p.417). But it is important to see that this inference is to an empirical claim. There is no question that the mathematical model can accommodate a structure such as Wright imagines, giving rise to a significant proportion of altruist individuals among the migrants. The question is whether or not this structure is representative of a natural population. Wright thinks that the existence of such traits (for group advantage at the expense of individual advantage) is so obvious that the question is not *whether* group selection can account for them, but *how* it does so (Ibid.). His model is therefore one possible explanation for the persistence of altruistic traits by group selection, but it remains to be seen if its structure is appropriately analogous to what is found in nature.

Wynne-Edwards (1962) argues that, indeed, Wright's structure can be found in nature, citing observations of "self-sterilizing behavior" that leads individuals not to breed even though other members of their species in the population are successfully breeding. And as Maynard Smith (1964) observes, if this "altruistic" population-regulating behavior is genetically determined, then only the "selfish" breeder genotype would ever survive. If fitness is measured by offspring, then the self-sterilizing individuals are clearly less fit. Nevertheless, they benefit their group as a whole. By limiting their breeding, their groups are less likely to outstrip their food supply and starve. Thus the "altruist" groups are more fit than the "selfish" groups.

To deal with Wynne-Edward's example, Maynard Smith offers his own mathematical model. He invites us to imagine a farmer's field populated by a species of mouse that breeds only in haystacks. Each year, when the haystacks are built, they are colonized by a male-female pair, who then reproduce. At harvest time, the stacks are destroyed, causing all of the mice to scatter back into the field as migrants until the following year when the haystacks are rebuilt, and the cycle repeats.

Unlike Wright's model, the total population here consists of only two phenotypes: selfish individuals, with genotypes AA or Aa, and altruists, of genotype aa. The selfish are individually

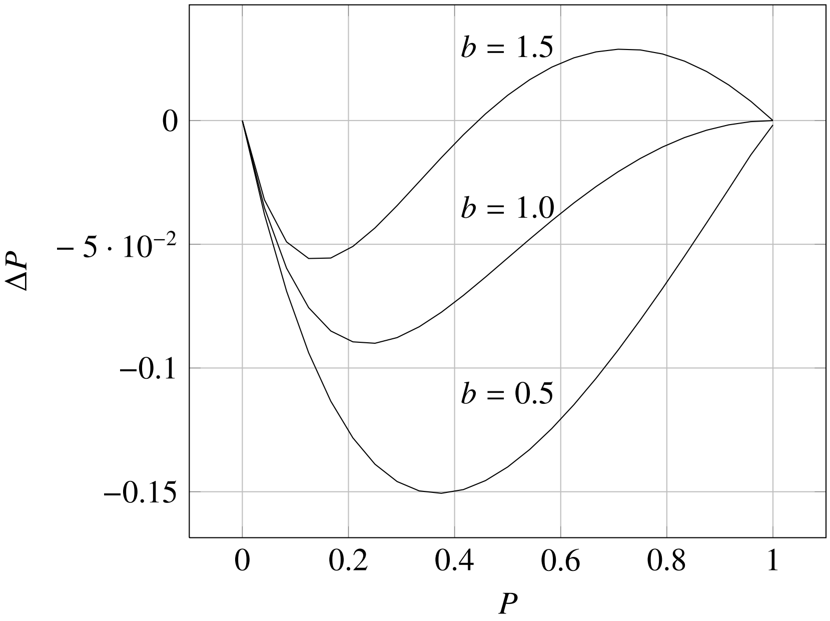

Figure 3.1: Maynard Smith's "haystack model", random mating

more fit, as they will always eliminate the altruist genotype in any mixed haystack; however, a pure altruist haystack will produce $(1 + b)$ times more offspring that a purely selfish haystack, and for all $b > 0$ will be more fit as a deme.

If we let P be the proportion of altruist haystacks in one year, then the proportion of altruist migrants is:

$$\frac{P(1+b)}{P(1+b)+1-P} = \frac{P(1+b)}{1+bP} \tag{3.4}$$

Assuming that all the mice find new haystacks and mate at random, the frequency of altruist haystacks the following year, P_1, will be:

$$P_1 = (\frac{P(1+b)}{1+bP})^2 \tag{3.5}$$

The altruist gene spreads whenever $P_1 - P = \Delta P > 0$. If we assume that the altruist mice are twice as productive as the selfish, that is, $b = 1$, then as Figure 3.1 shows, no matter what the initial proportion of the altruist genotypes, they will decline in frequency (i.e., ΔP is always below 0). This is consistent with Wright's earlier result.

Maynard Smith also considers cases in which the matings are not completely random. If we allow that some proportion, r, of the altruist mice will only mate with other altruists, then P_1 is redefined as:

$$P_1 = r(\frac{P(1+b)}{1+bP}) + (1-r)(\frac{P(1+b)}{1+bP})^2 \tag{3.6}$$

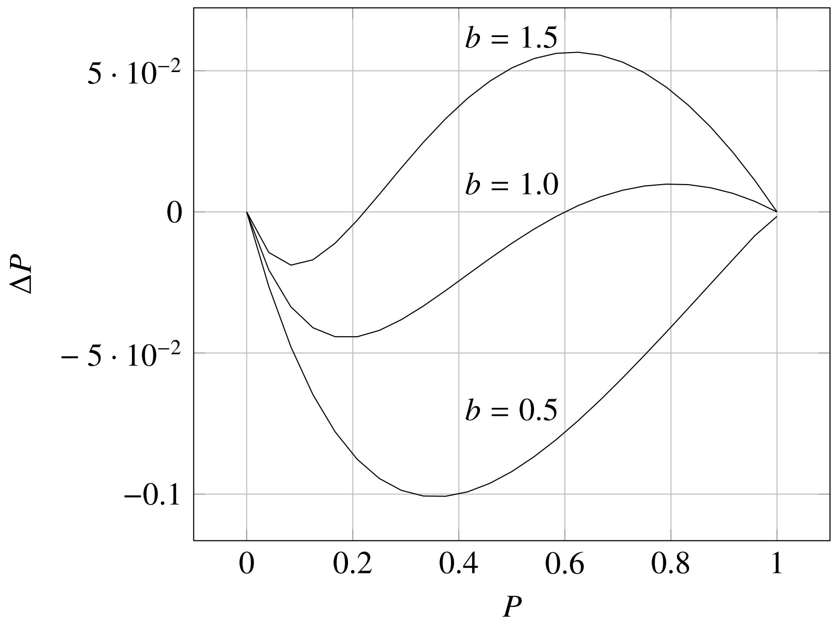

Figure 3.2: Maynard Smith's "haystack model", $r = 0.2$

The original model, with random mating, is thus a special limiting case of this new model when $r = 0$. Figures 2 and 3 illustrate how the dynamics of the system change with $r = 0.2$ and $r = 0.8$ respectively.

It is apparent from Figure 3.2, that with 20% non-random mating, the altruist genotype will spread more readily, provided that $P \approx 1$ and $b \approx 1$ or $b > 1$. But when $b = 0.5$, the result is little changed from the completely random mating case. Notice also that with 80% non-random mating, as shown in Figure 3.3, the altruist genotype spreads very readily. However, Maynard Smith objects that "the conclusion that timid or altruistic behavior can readily evolve if there is no interbreeding between groups means little, since it is unlikely that species are often divided into a large number of small and completely isolated groups" (Maynard Smith 1964, p.1146). In other words, the dynamics considered under the haystack model do admit the possibility of group selection, but the initial conditions and parameter values required for its efficacy are too extreme, on his view, to be explanatory for natural systems.

Wright and Maynard Smith provide an interesting contrast on this point. Both of their models show that group selection is possible under certain specifiable conditions: a combination of isolated (or semi-isolated) demes and a significant "altruistic benefit". Yet, they reach opposite conclusions about the explanatory relevance of the models for natural systems—a clear instance of Duhem's problem! Wright takes his model to represent observed adaptations and thereby explain how they can emerge from a certain population structure. Maynard Smith

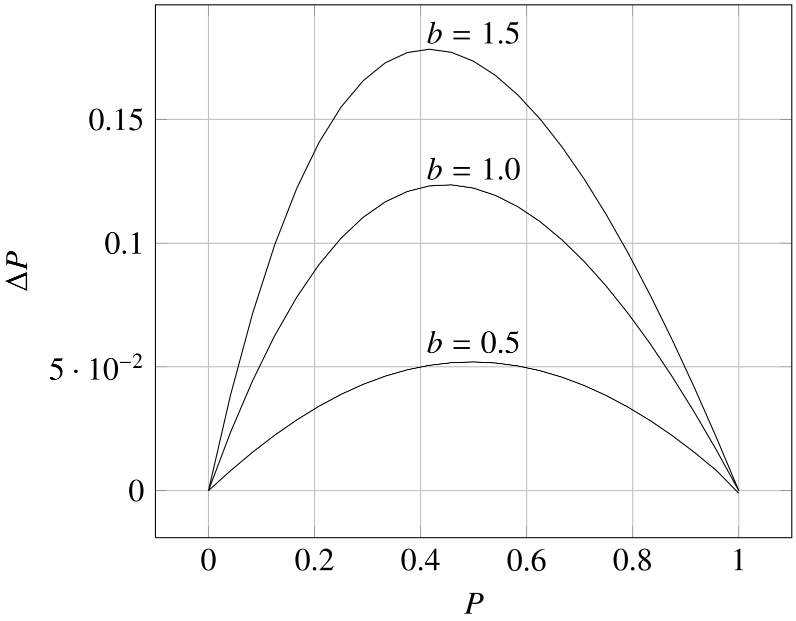

Figure 3.3: Maynard Smith's "haystack model", $r = 0.8$

takes his model to show that the conditions necessary for group selection are too extreme to be explanatory of natural populations.

As noted already, Wynne-Edwards sides with Wright, and in his response to Maynard Smith's model, writes that he is not skeptical at all about the natural possibility of small and/or isolated breeding groups. Wynne-Edwards attributes his disagreement with Maynard Smith as arising from the "differences in outlook and experience between a laboratory geneticist and a field ecologist" (in Maynard Smith 1964, p. 1147). He continues:

> Most ecologists would agree that the prerequisite of group selection that calls for a subdivided population structure is commonly and indeed normally found in animals. [Maynard Smith] says that the *Ortstreue* or return of migrant birds to their native locality would not bring it about; perhaps it is easier to see then in the case of the salmon or trout spawning in its natal tributary stream, where it more obviously becomes a member of a partially isolated breeding group (Ibid.).

But despite Wright's and Wynne-Edwards' explicit dissent, Maynard Smith's interpretation became the dominant one in biology. Five similar models were later elaborated from Maynard Smith's mathematical structure (cf. Levins 1970; Eshel 1972; Boorman and Levitt 1973; Levin and Kilmer 1974; Gilpin 1975), and each of these was taken to share the conclusion that although group selection was possible, it required extreme parameter values that were unlikely to be found in nature. By 1976, Maynard Smith is referring to this as the "orthodox view" in biology (Maynard Smith 1976, p.277), a sentiment evident to both Wade (1978) and D. S. Wilson (1983), and in part, responsible for each of their reviews.

In retrospect, it is odd that mathematical models could be taken to refute or restrict the possibilities of empirical evidence. Armed, as we now are, with a more systemic understanding of natural population structures, it may appear obvious that the promotion to orthodoxy of the conclusion against group selection's efficacy was hasty or ill-informed. It should have been revised as new empirical studies were conducted. Levin and Kilmer (1974), whose model is included amongst those taken to support the orthodox view, explicitly acknowledge this point in their discussion:

> . . . it is difficult to assign a significant evolutionary role to a mechanism which may operate only under very restrictive conditions. However, we see little utility in emphasizing this interpretation. In spite of the dearth of empirical estimates of the necessary parameters, there is already an abundance of literature taking negative stands on the role of interdemic selection. *Thus, for heuristic reasons, and because we believe that the rejection of a hypothesis ought to be empirically based*, we have elected to take a positive stand in our discussion of the role of interdemic selection. We point out that the demographic and genetic conditions necessary for the evolution of altruistic characters by interdemic selection may well occur in natural populations (Levin and Kilmer 1974, p.540, emphasis added).

Wynne-Edwards' remark about the "differences in outlook and experience between a laboratory geneticist and a field ecologist" is made more concrete by Levin and Kilmer's admission here. Despite the results of their own mathematical model, they are inclined to resist the negative interpretation because they "believe that the rejection of a hypothesis ought to be empirically based". This is likely an obvious point for the field ecologist, but not necessarily so for the laboratory geneticist or mathematician. It also provides us with a heuristic about how to interpret empirical results versus theoretical results:

> β_2: An empirical result has epistemic priority over a theoretical result for justifying explanatory claims about natural systems.

Whereas the orthodox interpretation of the mathematical group selection models depends upon the opposite heuristic:

> β_3: A theoretical result has epistemic priority over an empirical result for justifying explanatory claims about natural systems.

Adopting β_3 rather than β_2 illuminates one reason why Maynard Smith's conclusion is opposite Wright's or Levin and Kilmer's. Maynard Smith concludes that group selection is unlikely to be found in nature. But this is not a mathematical claim. Is it only inferred from the results of his mathematical model. So what justifies this inference? Do we have good reasons for accepting or doubting the applicability of β_3? In other words, do we have any meta-heuristics to guide us?

A conclusion from chapter 2, that robustness and mathematical stability are importantly distinct, is relevant here. Wright's and Maynard Smith's models both show that the "altruist" genotype will decline in frequency under certain conditions. This result is *mathematically stable* in two different senses:

1. It is invariant under certain perturbations to the initial conditions. In Maynard Smith's model, provided that $b < 1$, the initial frequency of the altruist gene did not change the result. In fact, the result was *globally* stable—since no initial frequency of the altruist (excluding 0 or 1) had an effect on the result. While this is not a proof of Lyapunov stability, it is evidence that such a proof might be possible.

2. It is invariant under certain perturbations to some parameter values and equations. For example, in Wright's model, there was both a benefit parameter, b, and a cost parameter, s. In Maynard Smith's model, there was no cost parameter. Yet, this change in the governing equations of the model did not change the result.

   The value of b could also be perturbed. In Wright's model, it was subject to the limitation of inequality 3.1, but independent of the change in gene frequency (equation 3.2). In Maynard Smith's first model (figure 3.1), b could take any value between 0 and 1 without changing the result.

   However, the result is not globally stable across parameters, since there were parameter values and equations under both models for which the altruist genotype would survive and thrive. Wright's introduction of the migrant parameter, m, and Maynard Smith's models with non-random mating were both more favorable to the altruist. These versions of the models also diminished the invariance across initial conditions, since the altruist's survival becomes more sensitive to its initial frequency.

   As above, this is no rigorous proof of structural stability, but it does provide some reasons to think that such a proof could be given.

These stabilities (and their limitations) are properties of the mathematical result (i.e., $\Delta q < 0$ for Wright or $\Delta P < 0$ for Maynard Smith). They tell us about relationships among the variables in the mathematical representation. But without some further evidence and justification, they do not tell us about relationships in the natural system. Whether group selection is efficacious is not ultimately a question about mathematical models, it is question about a real process in nature. Wright's and Maynard Smith's models show a range of possible conditions and structural assumptions that might allow for group selection to be efficacious. But this is entirely within the context of simplified mathematical representation. Without further work to either discount Wynne-Edwards' and other field biologists' observations or demonstrate how the simplifying assumptions of the mathematical model are well-grounded in empirical evidence (as Levin and Kilmer point out), the inference to group selection's inefficacy in natural systems remains tenuous at best. The later elaborations of Wright's and Maynard Smith's structures by Levins (1970), Eshel (1972), Boorman and Levitt (1973), Levin and Kilmer (1974) and Gilpin (1975) are therefore beside the essential point. The representational structure can be made as elaborate and sophisticated as one likes. Until the basic assumptions have been shown to be robust—reliable and unbiased for the natural systems in question—the inference from mathematical stability to explanations in nature remains unjustified. This is, in effect, to reject β_3 as a suitable heuristic.

3.2 Wade's Experiment and Analysis

Wade's (1976, 1977) experimental work finally takes those vital empirical steps, helping to explore the problem space between mathematical modeling and explanations of natural populations. Investigating group selection's efficacy in the laboratory with populations of the flour beetle, *Tribolium castaneum*, he sought to test the possibility of group selection resulting from the differential extinction and recolonization of populations.

The "group trait" chosen for selection was the number of adults in a deme. Since the number of adults in a deme is a property only of the deme (i.e., the relevant group for this test of group selection) and not of the individuals, if it can be selected for and shown to have an effect on subsequent generations, then this would provide evidence in favor of the efficacy of group selection, over and above the effects of collective individual selection, and hence, provide evidence to restrict the domain of β_1.

To test this, Wade took 192 beetle demes with 16 adults per deme and divided them into 4 different treatment arms of 48 demes each. At 37 days (enough time for an individual beetle to grow from an egg to a mature adult), a census of the number of adults in each deme was taken. The selection treatment was then applied. In treatment A, the deme with the largest number of adults was selected first and divided into as many new demes of 16 adults as possible. Then the deme with the second largest number of adults was likewise divided into new demes of 16, and so on until 48 new demes had been established.

In treatment B, the same general procedure was followed, except that the deme with the lowest number of adults was selected first and then divided into new demes of 16, then the second lowest, and so on. In treatment C, one deme of 16 adults was chosen at random from

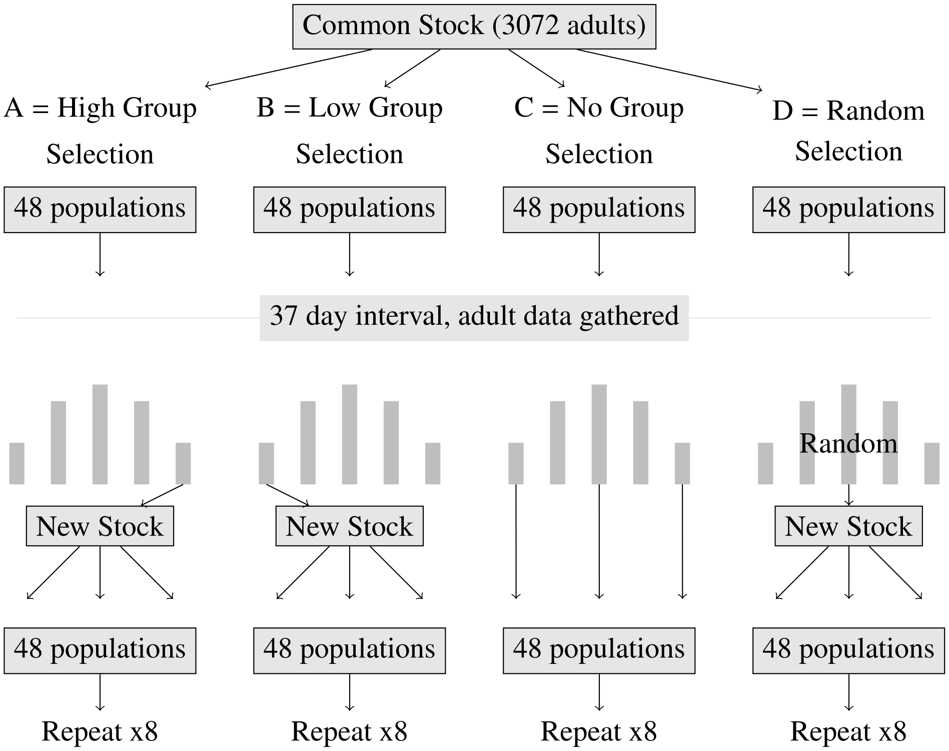

Figure 3.4: Wade's experimental group treatments

each of the 48 demes and then it alone was used to establish the new deme. This made treatment C the control, since only individual selection would be operating. In treatment D, the seeding demes were selected randomly before being divided into new demes of 16. The difference between C and D being that for C, *each* deme from amongst the 48 contributed 16 random adults to the next generation; for D, *one* deme was randomly selected and divided into new demes of 16, then another was randomly selected and divided, and so on until 48 new demes were established.

This selection process was then repeated 8 more times, making 9 generations in total (see figure 3.4). If group selection could not override the effects of individual selection, then one would expect treatments A, B, and D to be statistically similar to C; however, as early as generation 3, the mean number of adults in treatment A exceeded that of treatment C by more than 40 adults. Similarly, the mean number of adults in treatment B was 30 adults less than that of treatment C. These differences only grew more extreme. By generation 5, the mean number of adults in treatment A exceeded treatment C by over 100 adults (Wade 1977, p.139). Even treatment D began a slow, steady increase in mean number of adults versus the control in generation 4 (Ibid., p.141).

Wade takes his experiment to pose a serious challenge to the mathematical models of group selection and their support of the orthodox view. However, it is important to note that, at least in one way, their results are concordant. The mathematical models showed that highly isolated demes would be favorable to group selection, and indeed, the demes are completely isolated for Wade's experiment.

Nevertheless, as a follow up to his empirical work, Wade (1978) reviews the seven "traditional models" of group selection I cited above, as well as, five "intrademic models" (Wilson 1975, 1977; Gadgil 1975; Charnov and Krebs 1975; Matessi and Jayakar 1973, 1976; Cohen and Eshel 1976). The difference is that a traditional (interdemic) model assumes randomly mating local demes which then contribute migrants to a larger, population-wide mating pool after each generation; the intrademic models assume a single randomly mating deme whose members are organized into smaller "trait groups".

Despite a modification in mathematical structure between the traditional and intrademic models (cf. Wilson 1983), the general conclusion, unfavorable to group selection, remained the same. Wade investigates the assumptions made in each of the twelve models and finds that they all share at least three of these five:

1. The frequency of a single allele in a deme is capable of changing (i) the probability of survival of that deme, or (ii) the genetic contribution made to the next generation.

2. All demes contribute migrants to a migrant pool, which is then redistributed at random to fill vacant habitats.

3. The number of migrants committed to the pool by a deme is independent of its size.

4. The variance among demes is created primarily by genetic drift.

5. Group and individual selection are assumed to be operating in opposite directions with respect to the allele in question (Wade 1978, p.103).

The fifth assumption is an empirical hypothesis, which may be true or false depending upon the trait in question. Recall that Wright considered this kind of group selection to be just one possible means of its operation, and yet, its inclusion in all twelve of the models might suggest that it is the only means of (detectable) operation. Wade's experimental treatment A actually shows that for the group trait of deme size, group selection can work in the same direction as individual selection.

The other four assumptions, while still having empirical implications, are more straightforward idealizations—known falsehoods introduced by simplifying heuristics. Given that these assumptions were shared across all twelve of the models, we should expect shared biases, and a reduction (if not an outright failure) of robustness in the result.

It need not have been so. Wade illustrates how the mathematical models could have made some different assumptions. For example, instead of assuming a linear fitness landscape in which populations are genetically subdivided but experience a uniform selection environment (as is implied by the first assumption), the models might have constructed multi-peaked fitness environments, in which single genes or gene frequencies offer only temporary or local advantage (Ibid., p. 104-105).

The second assumption, that of the migrant pool, is perhaps the most problematic for group selection. Wade acknowledges that (at the time) "the colonization of new or vacant habitats [has] not been systematically studied in natural populations", nevertheless, some general features of natural migrations are well understood and should lead us to believe that a migrant pool is a very poor representation of the natural world. For example, it places no restrictions on how the particular individuals may colonize, and in effect, assumes no geographical barriers or behaviorally restricted movement patterns, which are both strongly believed to be central constituents of natural population structure (Ibid., p. 106). Wynne-Edwards' example of salmon returning to their natal stream is a perfect illustration of behaviorally restricted movement patterns in nature.

The migrant pool is a construct justified from a "mathematical point of view", Wade says, since (a) it sets the expected value of the gene frequency in newly colonized habitats equal to the frequency in the migrant pool and (b) its frequency is equal to the average frequency of the surviving populations (Ibid., p. 106). This simplifies the analysis of the problem, but effectively eliminates the possibility of detecting group selection effects in the model. In effect, the migrant pool mixes all the demes together. Insofar as new colonies are drawn from this pool, there may as well have been no demes at all. The third and fifth assumptions only further compound this bias, ensuring that any group advantage, measured in terms of numbers of offspring, is washed out at each generation (Ibid., pp. 108-109).

As an alternative, Wade proposes the construct of a "propagule pool", which weights the contribution to the mating and migration pool from each individual deme according to its size and does not allow all the migrants to interbreed. The result, he argues, is a more realistic representation of colonization, as larger groups will tend to produce more colonies. But he also acknowledges that most natural colonizations and migrations will fall somewhere in between a propagule and a migrant pool (Ibid., p. 109-110).

Wade's analysis of the group selection models, their assumptions, and his alternatives illuminates the importance of meta-heuristics to guide model construction. The mathematical group selection models and modelers can, by and large, be thought of as adopting one kind of research strategy; utilizing a particular set of heuristics for simplification and explanation. I argued in the previous section that the heuristic, β_3, permitting mathematical stability to justify natural explanations was faulty. Here we can understand Wade as, in part, arguing for the alternative, β_2. And we can also make one of the conditions for β_2's application explicit:

> B_1: For the study of group selection, apply β_2.

Indeed, Wade's critique of the mathematical models stems largely from their failure to make justified idealizations. That is, his suggested revisions are still idealizations—they simplify the description of the system—but do so in a way that he believes to be both more realistic and open to the empirical possibility of group selection. This strategy is made explicit in B_2, a consequence of which is the use of models whose assumptions are informed by empirical results.

The alternative method, which includes β_2, and is exemplified by Maynard Smith's haystack model and the subsequent models derived therefrom, is quite divorced from empirical research (remarkably so given the initial exchange between Maynard Smith and Wynne-Edwards). It preserves certain simplifying constructs, presumably for their analytical tractability, at the cost of overlooking the complexity of natural population structures.

But we can go still deeper than this to discuss more of the particular heuristics of model construction responsible for the bias in the group selection models. Take, for example, the heuristic (identified in Wimsatt 1982) of *context simplification*:

> β_4: When constructing a model, we should simplify the description of the environment before simplifying the description of the system.

Using β_4 is often a very effective way of describing, analyzing, and explaining a phenomenon; nevertheless, by ignoring most or all of the features of a system's environment, it will make it difficult, if not impossible, for the model to detect interactions between the environment and system variables. The characteristic bias of this heuristic is perhaps most readily apparent in the assumption of the migrant pool. By assuming effectively no group barriers in migration and colonization, the group selection models were biased against the efficacy of group selection, which critically depends on the existence of isolated breeding groups.

Descriptive localization is another of Wimsatt's heuristics that applies here:

> β_5: Describe a relational property as if it were monadic, or a lower order relational property

Wimsatt's illustrative example of β_5 is the concept of "fitness", conceptualized as a property of a pheno- or genotype, rather than a relation between the phenotype and the environment (Wimsatt 1982, p.174). Assumption 1, identified by Wade above, is an obvious result of this heuristic.

Applications of β_4 and β_5 introduce a similar bias. They both ignore interactions between variables and the environment. Given that bias, there will be a whole class of problems for which they are ill-suited. As Wimsatt (1982) identifies, they are particularly ill-suited to problems that overlap multiple levels of organization. Wherever multiple levels of organization are relevant to a problem, simplifying away the possibility of detecting effects among those levels is going to be disadvantageous, and likely lead to errors. Indeed, group selection is just such a multi-level phenomenon. In order to observe group selection effects in a model, it must be possible, within the model, to distinguish between the properties of the individuals and the properties of the groups.

Therefore, we can specify another, more general meta-heuristic about the appropriate use of β_4 and β_5:

> B_2: For problems that overlap or involve multiple levels of organization, avoid or limit the use of β_4 and β_5.

Wade's choice of studying the group trait of population size rather than the individual trait of altruism, so often chosen for the mathematical models, speaks to an implicit recognition of B_2. Deme size is not a property of the individuals, and therefore, the fact that it can be selected for shows that selection can be operative at more than one level of organization. In contrast, the fifth shared assumption Wade identifies, that group and individual selection work in opposite directions, does not allow for the possibility that individual and group selection may both be explanatory, simply at different levels of organization.[2] Understanding this kind of general relationship between a heuristic, the idealizations produced by its application, and the bias introduced is a crucial part of the content informing the second-order judgments of good sense.

Further complicating matters, we should observe that applying context simplification will likely be necessary, to some extent, in every scientific model. There is thus a continual need to explicate the meta-heuristics, like B_2, in ever greater detail, articulating exactly why and when certain simplifications, in kind and degree, are appropriate. This is what I mean by distinguishing "benign" idealizations, which are useful, appropriate, and perhaps even necessary for understanding a problem; and "malignant" idealizations which are inappropriate, prone to unnecessary error, and conceal bias. Comparing the idealizations of the migrant pool with Wade's alternative—the propagule pool—it is clear that the latter is a less simplified description of the environment, and therefore, insofar as B_1 and B_2 constitute good sense here, we can judge that the propagule pool is more benign.

[2] The prevalence of that view is, in part, evidence of Williams' heuristic β_1. If individual selection is sufficient for explanation, then group selection operating in the same direction is just assumed to be superfluous.

But it is important to see that just as no single model can establish robustness, neither can it be responsible for the failures of robustness or presence of malignant idealizations. Robustness as a strategy prescribes that we perturb heuristics across the space of sampled models. Taken individually, the five assumptions that Wade identifies are not particularly problematic. All models must employ some idealizations, and there are surely epistemic conditions under which each of these would be useful. Maynard Smith's haystack model, as idealized and unrealistic as it is, is not a bad place to begin. The problem arises because the subsequent models did not adequately perturb their heuristics (i.e., failed to recognize E_3). They all ended up with essentially the same idealizations, and therefore, there was no check against the possibility of bias.

Finally, connecting these ideas back to those in the first chapter, we should note how Wade's analysis is an example of good sense. His experimental result conflicted with a theoretical prediction. So what does he do? The fact that four of the five assumptions Wade identifies as problematic are idealizations immediately complicates the usual story accompanying discussion of Duhem's problem and underdetermined judgment. In the face of recalcitrant evidence, Wade is not rejecting one hypothesis which was thought to be true or false. He is revising an interrelated group of assumptions, most of which were already known to be false, but which were employed heuristically for the purpose of simplification.

3.3 Explanatory Pathways and Group Selection's Problem Space

Griesemer and Wade (1988) propose the "pathways of explanation", a more principled tool to distinguish between robustness and pseudo-robustness in biological explanations. Shown in figure 3.6, the pathways essentially amount to a detailed map of the problem space for biological explanation. On their picture, there are three kinds of systems and investigations that contribute to our understanding: natural systems and field work; laboratory systems and experimental models; and mathematical systems and mathematical models.

As we know, heuristics play a role at each "conceptual stage": determining the content of the work plan or model specification; setting the experimental protocol or simulation parameters; etc. The heuristic β_2, emphasizing the priority of empirical work over mathematical work, is already represented here, since the "pathway" to explaining "agents in the natural systems" requires the use of laboratory and field studies (rather than mathematical models). Griesemer and Wade argue further for this meta-heuristic, noting that the more the system under study can

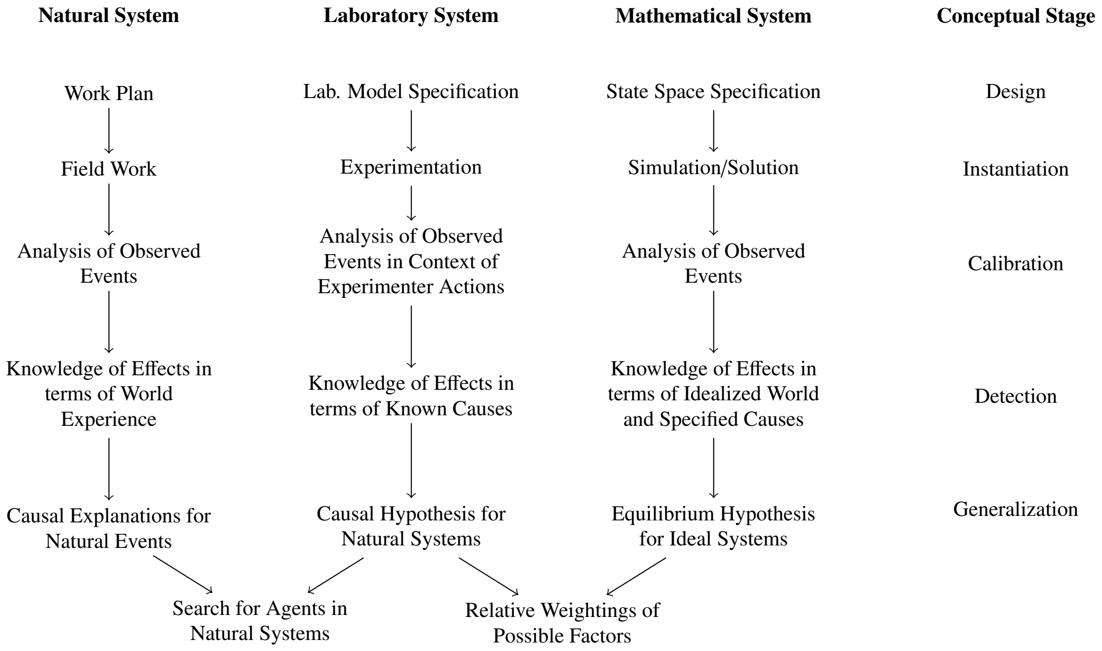
Natural System
Laboratory System
Mathematical System
Conceptual Stage
Work Plan
Lab. Model Specification
State Space Specification
Design
Field Work
Experimentation
Simulation/Solution
Instantiation
Analysis of Observed Events
Analysis of Observed Events in Context of Experimenter Actions
Analysis of Observed Events
Calibration
Knowledge of Effects in terms of World Experience
Knowledge of Effects in terms of Known Causes
Knowledge of Effects in terms of Idealized World and Specified Causes
Detection
Causal Explanations for Natural Events
Causal Hypothesis for Natural Systems
Equilibrium Hypothesis for Ideal Systems
Generalization
Search for Agents in Natural Systems
Relative Weightings of Possible Factors

Figure 3.5: Griesemer and Wade's pathways of explanation

be thought of as a plausible subclass of natural systems, the stronger the explanatory inference.[3] They write:

> Because laboratory models are actual material causal structures, the pathway of analogical inference is from *actual* (i.e., real existent, competent, responsible) causes to actual artificial effects, to matching natural effects, to claimed actual natural causes. ... With mathematical models, the path is from mathematically *possible* representations of natural causes, to actual mathematical effects, to actual or simulated natural effects, to claimed actual natural causes.
>
> Laboratory systems are easily viewed as a subclass of natural systems in which scientists are important agents. ... Mathematical systems can only be viewed as a subclass of natural systems in a world too Platonic for most scientists to accept (Griesemer and Wade 1988, pp.77-78).

The discussion in §3.1 makes it clear that the analogy from the mathematical group selection models to the natural system was too weak; demonstrating stability properties, but lacking

[3] There is a similarity between Griesemer and Wade's claim about biological models and natural systems and the relationship between pragmatic trial designs and generalizability (i.e., external validity) of trial results in clinical research (cf. §2.2.1). In much the same way that a more pragmatic trial provides better evidence of treatment effectiveness in clinical practice, because it is a more plausible surrogate of the target environment than a highly explanatory study, so too is a field study or laboratory model a better surrogate for the conditions in natural systems.

in robustness and empirical support. But this is not to say that the mathematical models could be of no use to the study of group selection. Rather, the conclusion Griesemer and Wade want to draw is that mathematical models are insufficient *on their own* to ground robust, explanatory claims about natural systems. Where their assumptions have been verified in laboratory models or have already been shown to be robust across natural systems, mathematical models can be explanatorily adequate. It is simply that the "pathway" of explanation from their results to natural systems is different.

The most important of Griesemer and Wade's pathways for understanding the group selection debate are the second and third columns, corresponding to laboratory and mathematical systems, respectively. On their account, the proper explanatory inference to make from a mathematical system involves the relative weightings of parameters. As we saw, Maynard Smith's analysis of his haystack model does draw these inferences. The error emerges when he goes too far and makes the unwarranted claim that the parameter weightings necessary for group selection are rarely found in nature. Following the pathways, we can immediately see that for such a claim, a laboratory model or field work would is needed.

Putting the lessons from the pathways together with the preliminary picture of the modeling problem space discussed in the last chapter, we can now articulate a more complete problem space for group selection. The pathways articulate the various kinds of investigations (the columns) and the epistemic steps (the rows) toward the ultimate explanation of some natural phenomenon. The modeling space of figure 3.6 gives us a way to illustrate the diversity of heuristics and modeling approaches, and hence, to judge the robustness of the research program.

As in chapter 2, we can begin with Levins' (1966, 1993) modeling dimensions: realism, precision, generality, manageability, and understandability; and then overlay multiple plots, each representing a different kind of model or experiment, in order to discuss the diversity of explorations. Let the solid line in the figure correspond roughly to Maynard Smith's family of mathematical group selection models. They are maximally manageable and precise. There are no random fluctuations, no need to run multiple simulations, and they provide definite quantitative results, dependent only upon the initial conditions and parameter values. They are also highly general, since they are really just mathematical structures thought to apply to many different kinds of systems. Judging from Wade's analysis, we can say that they are not realistic, and from Levin and Kilmer's divergent interpretation of the result, we might add that they are not straightforwardly understandable.

The dotted line corresponds roughly to Wade's experiment. As a laboratory model, it is surely less precise than the mathematical models; also less manageable. It is less general, since it was run with a particular species of flour beetle, and the results may depend upon

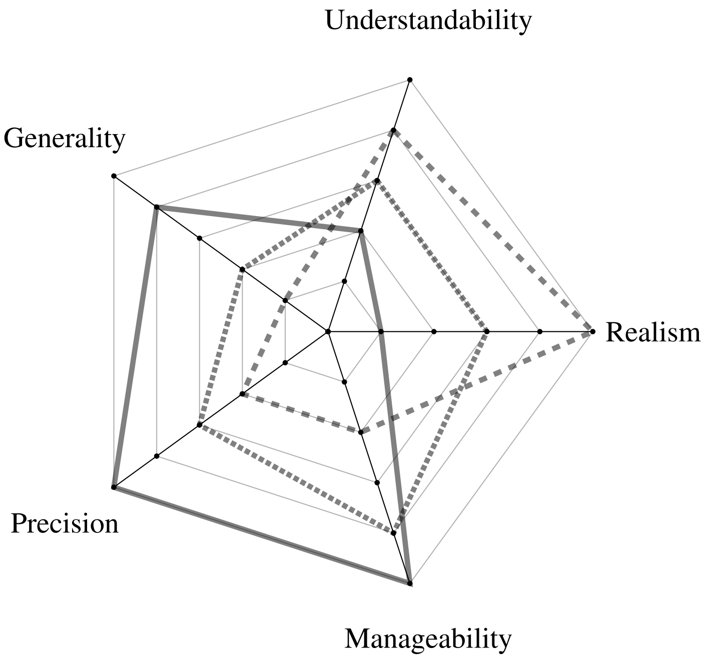

Figure 3.6: Problem space for group selection studies

the breeding behavior of that particular organism. This is not to say that Wade's conclusions cannot be generalized to other species under similar conditions, but this inferential step requires some more justification as to why the perturbation from one species to another would not change the conclusion. As a result, the interpretation of the experiment is also somewhat less understandable. One cannot simply reconstruct and check the laboratory model, as one might a mathematical model. We rely, in part, upon what Wade tells us in his published methods, results, and discussion sections.

But importantly, the laboratory model is more realistic. It is investigating group selection with real populations of living organisms. The extent to which natural populations are similarly organized is still a relevant factor; nevertheless, there is no question that Wade's flour beetles are more like flour beetles in their natural habitats than Maynard Smith's mathematical field mice are like real field mice in theirs.

These two different kinds of models (mathematical and laboratory), and the different areas they occupy on the graph, begin to illuminate some robustness considerations. Although no model (or kind of model) can fill the entire space, a robust program of research will include many different kinds of models, which when taken together, can explore much of the space. And indeed, the meta-heuristic for robustness (cf. Γ_1 in ch.2) prescribes that a program of research ought to try and explore the space. The graph helps to makes clear how the mathematical models alone do not cover a large portion of the possibilities, and in particular, neglect the critical region associated with realism.

The group selection case is interesting in this regard, since the result of Wade's laboratory model was discordant with the mathematical models. With the aid of the graph, one can now see that the discordance corresponds to explorations in different regions of modeling space. The more realistic regions are favorable to group selection's efficacy, the less realistic are less favorable. But contrary to Stegenga's (2009) view discussed in the last chapter, this situation does not leave us helpless. Wade takes this failure of robustness to be a reason to re-evaluate the structure of the mathematical models. In effect, his recommendation is to move the mathematical models into a more realistic region of problem space.

Finally, the dashed line in figure 3.6 could roughly correspond to the often cited field study of the *myxoma* virus in rabbit populations (Lewontin 1970; Fenner and Ratcliffe 1965). To describe the case briefly: an avirulent strain of the disease was observed to survive and displace its more virulent relative. Individual selection would seem to favor the virulent strain, with its higher growth rate. However, these tend to kill their hosts quickly, driving the entire disease population within the host extinct. An avirulent strain has a slower growth rate, kills its host less quickly, and allows the population to survive long enough to infect other hosts—provided, of course, that the populations in live hosts are more successful in spreading their infection than the populations in dead hosts. The avirulent strain, while less individually fit compared to its more virulent relative, nevertheless increases the survival rate and fitness of the entire disease group (Wilson 1989, p. 68).

As a field study, this result is going to be less controlled, less precise, and less general. Nevertheless, it is maximally realistic, since it is simply measurement and observation of a natural system. Its understandability raises some further questions of interpretation. D. S. Wilson (1989), for example, takes the study to display an obvious case of group selection. R. Wilson (2004) disagrees with this interpretation of the study, and instead takes the myxoma case to show that the very notion of "levels of selection" is ill-conceived. He points out that the *group* on which group selection is supposed to act to bring about avirulence is not as obvious as it might seem:

> . . . if rabbits constitute a group of viruses, then surely (given contagious infection) a *hutch* of rabbits does as well, as does the *local deme* of rabbit hutches. And given the differential transmissability of viruses located on different parts of the rabbit, all those viruses on the rabbit's head, or its ear, or around its eyes, also constitute groups of viruses . . . for although rabbit-bound viruses are not an evolutionary arbitrary group, they are far from a unique such group, and the focus just on rabbit-bound viruses as the object of group selection seems either arbitrary or in need of further justification (Wilson 2004, p.395, original emphasis).

R. Wilson's dissent about this case, and whether or not it supports arguments for group selection, illustrates two points: First, how the myxoma studies are open to multiple interpretations, and therefore, warrant the classification of reduced understandability (cf. table 2.2); and second, that the controversy of levels of selection in biology is still being re-interpreted. Part of my argument here is that a well-articulated problem space, which I am only beginning to describe, can sharpen the debate. If R. Wilson is correct, and the myxoma case is not evidence either for or against the efficacy of group selection, then its plot can be eliminated from the robustness graph in figure 3.6, leaving the overall research program with a clearly under-explored region.

This illuminates a prospective advantage of representing the problem space as I have done it here. A potential biologist, interested in testing the robustness of earlier results, can literally see the unexplored areas of research, and work to construct models and studies to occupy these regions. This also makes the methodological weight of Wade's alternative idealizations (e.g. a multi-peaked fitness environment or a propagule pool) more explicit. These are assumptions that would push knowledge further along the realism dimension. Insofar as that region remains unexplored, the inferences to natural explanations are weakened.

Chapter Summary

Although contemporary debate about the units and levels of selection has largely moved on from Maynard Smith's haystack model and Wade's experimental result, the divergence in the understanding and interpretation of their work reveals methodological lessons that have not been recognized. Wade, Wimsatt, and Griesemer and Wade all suggest insightful meta-heuristics in their analyses. Yet, without an epistemic framework for linking these and other methodological prescriptions together, it can be difficult to move beyond a polarized picture of differing scientific judgments, where the philosophical conclusion seems only to be that some thinkers were prescient or had "insight" that others did not. By making the underlying heuristics and meta-heuristics explicit, we can bring fundamental differences among research strategies into sharp relief.

The orthodox understanding of Williams' heuristic, β_1, as prohibiting group selective explanations, is clearly contrary to good sense. The domain of β_1 is more limited and must be held to empirical standards. Similarly, Maynard Smith's inference from his mathematical result can be recognized as unjustified. But this should not undermine the epistemic value of what he does actually show. His model, and those that adopted his structure, were able to demonstrate a remarkably stable result. They did not, however, establish a *robust* result. Just as Wade's experimental work is not, by itself, a robust demonstration of the opposite result.

I have argued that each of these investigations contributes to a greater understanding of the problem space, and that the persistent disagreement and failures of robustness are signs that we still need to perturb our heuristics toward new kinds of investigations. The problem spaces that I describe help to make this prescriptive judgment more explicit. A new biological model that is only capable of extending a result's claim to mathematical stability is clearly less valuable than a new experiment which evaluates the weightings of parameter values from a previously constructed mathematical model. The latter answers an important question about robustness, and in this sense, it contributes information that is of greater value to the overall problem space. But perhaps what was most needed for the group selection models was extensive field work to identify and categorize different population and migration structures. This information could have then been used to calibrate the realism dimension.

In closing, I note that while it is not often discussed as such, the group selection controversy is a textbook case of underdetermination. Empirical and theoretical results conflict, necessitating that the scientists come to a judgment about how to proceed. Focusing on the *problem* of underdetermination, as philosophers of science often do, perhaps makes this case unremarkable: It is simply one of a vast number of underdetermined judgments in the history of science. Decisions were made, theoretical wrinkles ironed out, and eventually scientists come to agree (even if they later realize their agreement may have been overhasty). But this usual philosopher's story overlooks too many of the important methodological considerations. As we saw here, hidden bias in modeling, failures of robustness, and drawing faulty explanatory inferences all inhibited (at least for a time) positive scientific development. These are all general problems faced in science, and the prescriptive guidelines I have drawn out here (e.g., the heuristic β_2 and the meta-heuristic B_2) go well beyond the particular case. It is thus by focusing instead on the *solutions* to underdetermination, articulating the relevant meta-heuristics, that the story becomes more philosophically fruitful.

Chapter 4

Assay Sensitivity and the Ethics of Placebo-Controlled Clinical Trials

While the previous chapters have focused on historical case-studies, looking back on extant work in science and philosophy of science in order to extract what we can about good sense, in this chapter, the question will be how to move forward in an ongoing debate. What is the nature of good sense for clinical trials such that it might require us to use a placebo control? And what are the relevant ethical constraints that might come into conflict with the epistemology? Both of these questions circle the concept of assay sensitivity, and so it is with that concept that we begin.

In February of 2010, the World Medical Association hosted an international symposium on the ethics of placebo controls in clinical trials. Despite years of debate, ethicists, clinical trialists, and policy makers remain divided over the ethical acceptability of using placebos in research when a proven, effective treatment is available. The protracted nature of this problem is due, at least in part, to a perceived conflict between the opposing demands placed on clinical research by ethics and science. Ethical standards demand that no patient enrolled in a trial should be deprived of competent medical care, and therefore, when a proven, effective treatment exists, it must be used as the active control. However, a scientifically valid trial, it is argued, must possess "assay sensitivity", and without using a placebo control, there can be no guarantee that it is.

There are a few definitions of assay sensitivity in the clinical research literature (cf. Hwang and Morikawa 1999, p.1208; ICH E10 2000, p.7; Temple and Ellenberg 2000, p.457; Gelfand et al. 2006, p. 944). The one that I will be concerned with here is that used by Robert Temple and Susan Ellenberg's (2000) article in the *Annals of Internal Medicine*, since it is their work that most profoundly shifted debate on the subject. They define assay sensitivity as the prospective "ability of a clinical trial to distinguish an effective from an ineffective treatment".

Prima facie, this would simply seem to be another way to describe a well-designed experiment. A well-designed experiment, whatever else it may do, has the ability to confirm or disconfirm an hypothesis. In the case of clinical trials, the hypothesis under consideration is the effectiveness of a treatment; and assay sensitivity seems to say only that a trial is capable of confirming or disconfirming this hypothesis. An essential property for an experiment to have, but not an obviously controversial or confusing one.

Unfortunately, controversy and confusion abound. Temple and Ellenberg (2000) argue that an assessment of a trial's assay sensitivity largely depends upon the control treatment used in its design. They contrast the supposed epistemic merits of placebo-controlled superiority trials, which test whether or not a new treatment is superior to placebo ("PCTs"), with active-controlled equivalency trials, which test whether or not a new treatment is as good as, or no worse than, a standard therapy ("ACETs"). Assay sensitivity, they claim, is assured with PCTs but not with ACETs.[1]

A closer inspection of Temple and Ellenberg's argument reveals that a number of important philosophical insights have been overlooked. The first half of this essay will thus be devoted to conceptual tidying. After providing some history and background for the question of control method (§1), I will then analyze the claims made about assay sensitivity and the argument for the methodological superiority of PCTs (§2). I argue that the supposed consequences of assay sensitivity and the seeming epistemic advantages of PCTs are entirely illusory; the result of (i) conflating biological efficacy and clinical effectiveness; (ii) oversimplifying the epistemology and ethics of placebos; (iii) assuming an implausibly radical form of underdetermination; and (iv) conflating the analysis of a *trial-as-designed* with a *trial-as-executed*. Once these points are clarified, the supposed conflict between science and ethics dissolves.

Having thus rejected the received understanding of assay sensitivity, the second half of the essay develops an alternative account of how the concept should be understood in the context of a clinical research program. In §3, I return to the straightforward reading of assay sensitivity, as meaning a prospective judgment that a trial is well-designed; a judgment appropriate to the trial-as-designed context. I then elucidate how the meaning of "well-designed" can be made more explicit by combining insights from the previous chapters on modeling in the philosophy of science and methodological tools in the clinical trials literature.

Finally, in §4, I sketch an analytical framework for the trial-as-executed context. As underdetermination makes clear, no trial, however well-designed, is capable, alone, of providing sufficient evidence of efficacy or effectiveness. What is needed, I argue, is a series of trials,

[1] Since the time of Temple and Ellenberg's writing, a third kind of trial, called a "non-inferiority study" has effectively replaced the equivalency design. A non-inferiority study tests whether or not a new treatment is better or no worse than (by some clinically significant margin) a standard treatment.

whose various designs are perturbed to optimally contribute toward demonstrating a robust pattern. Putting the tools of the previous section together, I offer an account of how a clinical research program could be structured to satisfy both the ethical and epistemological needs of clinical research.

4.1 Background

In 1963, Sir Austin Bradford Hill, one of the pioneers of randomized clinical trials ("RCTs"), wrote:

> Is it ethical to use a placebo? The answer to this question will depend, I suggest, upon whether there is already available an orthodox treatment of proved or accepted value. If there is such an orthodox treatment the question will hardly arise, for the doctor will wish to know whether a new treatment is more, or less, effective than the old, not that it is more effective than nothing (Hill 1963).

This short passage contains the seeds for much of the current controversy surrounding assay sensitivity. Implicit in Hill's assessment is an assumption about the kinds of questions that clinical trials should answer. For him, that question is a practical one (what "the doctor will wish to know") about the more (or most) effective treatment for a condition. This is distinct from other questions about a treatment's biological effects, its safety, or its potency.

There is also a more fundamental epistemological question about what constitutes "an orthodox treatment of proved or accepted value". Is it enough that an orthodox treatment is "accepted" by the medical community or must it have been evaluated in a PCT before it can be considered truly effective? As we will now see, these methodological issues—What question does a clinical trial seek to address? What sort of evidence is acceptable in medicine?—re-emerge time and again.

4.1.1 The spectrum of trial questions and trial designs

A distinction between questions of biological *efficacy*, on the one hand, and clinical *effectiveness*, on the other, is canonical in some methodological texts, but it is not a distinction that everyone recognizes or takes seriously. Friedman et al., for example, describe it thus: "[Efficacy] refers to what the intervention accomplishes in an ideal setting; [effectiveness] to what it accomplishes in actual practice, taking into account incomplete compliance to protocol" (Friedman et. al 1998, p.3). Unfortunately, drawing the distinction in this way is not as clear as it could be. While it does tie the investigation's question to certain aspects of the trial design (i.e., stricter setting and protocol tests efficacy; “looser” or more flexible setting and protocol tests effectiveness), it is not explicit about what these properties are.

I understand the distinction in a slightly different, albeit related, way: Efficacy refers to an intervention’s biological activity. What does the intervention do? What measurable effects does it have on the patient? How does it effect the target condition? These and other questions about dose response, safety, purity, and potency all fall under the category of efficacy questions. Consistent with Friedman et al.’s definition, these are all questions best investigated under laboratory conditions.

Effectiveness is as Friedman et al. describe it: An intervention’s performance in (conditions similar to) clinical practice. Do patients assigned to this intervention improve? Do patients improve more on this intervention than others? Is it more cost-effective than alternative courses of treatment? These and other practical questions about the relative merits of an intervention fall under the category of effectiveness questions.

Drawn in this way, the distinction is clear. Efficacy refers to biological *properties or activity*; effectiveness to *performance in clinical practice*. For example, showing that a cancer therapy shrinks tumors (in the study population) can sufficiently answer an efficacy question (e.g., “Does this drug have an effect on tumors?”). To sufficiently answer an effectiveness question (e.g., “Will this drug improve cancer patient survival?”), the therapy must be shown to improve the outcomes of patients in populations and conditions similar to those found in actual clinical practice.

A related contrast between the kinds of questions a clinical trial seeks to answer also appears in Schwartz and Lellouch (1967), as the difference between *explanatory* and *pragmatic* “attitudes in therapeutical trials”. They present the distinction as one between kinds of research goals: A pragmatic trial is aimed at reaching a decision about the better of two treatments for clinical practice; an explanatory trial is aimed at increasing our understanding of a treatment’s biological effect (Ibid., p.638). The natural conceptual relationship here is that pragmatic trials ask questions about effectiveness; explanatory trials ask questions about efficacy. But we should not be fooled into thinking that these categories of questions and aims are mutually exclusive. Thorpe et al. (2009)’s observation that most trials are actually a mixture of explanatory and pragmatic assumptions, and their PRECIS tool representing the spectrum of trial assumptions, is important to bear in mind (cf. §2.2.1).

Employing this vocabulary, we can say that Hill’s original perspective is that of a pragmatic-minded trialist. If we seek to answer a question about which of two treatments is better in practice, then a more pragmatically oriented trial, comparing a new treatment with the current standard therapy, is the best option. This would be a trial aimed at answering a pragmatic question about the relative effectiveness of a new treatment.

But this is certainly not the only kind of question worth asking. Regulatory agencies, for example, may be more interested in a treatment's efficacy. Indeed, the FDA's procedure for licensing new drugs calls for "data derived from nonclinical laboratory and clinical studies which demonstrate that the manufactured product meets prescribed requirements of safety, purity, and potency" (U.S. Code of Federal Regulations Title 21, §601.2). An explanatory trial, comparing a new treatment with a placebo, may be sufficient to show "safety, purity, and potency".[2]

The relationship between the orientation, or "attitude", of a trial's design and the question it seeks to answer is precisely that which PRECIS aims to make explicit. In effect, selecting the question to investigate also determines the choice of control. Pragmatic questions about effectiveness typically call for active controls to determine whether or not the new treatment is better, or no worse, than the standard. Explanatory questions about efficacy, on the other hand, may call for placebo controls to more precisely determine the magnitude of a treatment's effect.

4.1.2 Clinical equipoise and proven, effective therapy

The epistemic goal of a trial is only one part of the issue. The choice of control is also central to a debate about the duty of care in clinical research, which prohibits a physician from knowingly providing patients with inferior treatment. As Freedman puts it:

> In the simplest model, testing a new treatment *B* on a defined patient population *P* for which the current accepted treatment is *A*, it is necessary that the clinical investigator be in a state of genuine uncertainty regarding the comparative merits of treatments *A* and *B* for population *P*. If a physician knows that these treatments are not equivalent, ethics requires that the superior treatment be recommended (Freedman 1987, p.141).

Given such a duty, we might wonder: How is it that a physician could ever ethically enroll a patient into a clinical trial, exposing them to a new, potentially dangerous, and possibly inferior

[2] The FDA shares with Temple and Ellenberg and ICH E10 the troubling pattern, which we discuss below, of conflating "effectiveness" and "efficacy". For example, on the FDA's website, their home page devoted to "Centers and Offices" describes their responsibilities thus: "The FDA is responsible for protecting the public health by assuring the safety, *efficacy*, and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation," (USFDA, http://www.fda.gov/AboutFDA/CentersOffices/default.htm, Mar. 2010, emphasis added) whereas the page for "FDA Fundamentals" describes their responsibilities as "protecting the public health by assuring the safety, *effectiveness*, and security of human and veterinary drugs, vaccines and other biological products, medical devices, our nation's food supply, cosmetics, dietary supplements, and products that give off radiation" (USFDA, http://www.fda.gov/AboutFDA/Basics/ucm192695.htm, Mar. 2010, emphasis added). These two descriptions are almost identical except for the change from "efficacy" to "effectiveness". While this makes it difficult to determine exactly what the FDA considers its mandate to be, it makes it obvious that they do not appreciate the importance of the effectiveness/efficacy distinction.

treatment? As Worrall (2008) observes:

> Indeed, in the case of placebo-controlled trials, it is hoped and generally expected that the experimental treatment will outperform the placebo. How is this compatible with laudable sentiments that are taken to govern the practice of all physicians, such as "the health of my patient will be my first consideration" (Worrall 2008, p.423).

Freedman's concept of clinical equipoise—a state of honest, professional disagreement amongst the expert medical community regarding the therapeutic merits of the arms in a trial—offers a compelling answer. Such a state of disagreement depends on each arm of a trial being consistent with competent medical care. So long as this is the case, patient participation is consistent with the physician's fiduciary responsibility.

Importantly, the requirement of clinical equipoise is not waived for a PCT. It demands that there exist genuine uncertainty amongst the expert medical community about the net therapeutic advantage of a new treatment versus placebo (Freedman 1990). When standard therapy exists and is widely available, the requirement of clinical equipoise generally makes a PCT unethical, since patients in the control arm are receiving less than competent medical care.

Yet, as Worrall also observes, certain "hard-line" defenders of evidence-based medicine might object that until an intervention has been evaluated in a PCT, physicians do not have any *objective* evidence of its efficacy or effectiveness. A treatment may be the "accepted" standard of care, with physicians swearing to its effectiveness, but this is not a sufficient scientific justification for its use as an active control. Worrall summarizes the position that objective evidence is only provided in an RCT thus:

> So long as there is no objective evidence that one or the other treatment is inferior, a researcher's personal convictions are irrelevant. Or rather, those personal convictions ought to be critically examined and replaced by views that are in line with the objective evidential situation. It may prove difficult psychologically, but surely clinicians ought to ask themselves where any convictions about the greater effectiveness of a treatment come from, and, if the answer is "not from properly analyzed evidence", then they ought to modify those convictions. *This critical process might reveal that—despite their initial position with its associated but ill-founded convictions—they were objectively in equipoise because they had no really telling evidence that the control treatment was inferior; and hence that no ethical issues in fact arise about the trial* (Worrall 2008, p.424, emphasis added).

Essentially, the hard-line idea is that prior to evaluation in an RCT, or even more specifically, a PCT, the medical community *must* be in a state of equipoise, simply because there can be no real evidence of effectiveness. But Worrall, correctly, rejects this view, pointing out that there are great number of medical interventions that have never been and never will be tested in an RCT:

> For all the fact that evidence-based medicine advocates can point to *some* treatments (grommets for glue ear, and suppression of ventricular ectopic beats are favorite examples) that had been generally accepted in medical practice but then "proved" to be ineffective when subjected to trials, this surely cannot be true for *all* accepted but non-RCT-tested treatments. Surely there is strong evidence in favor of the effectiveness of some such treatments (such as appendectomy for acute appendicitis)—strong evidence that cannot, then, have been delivered by an RCT (Ibid., p.425).

Worrall ultimately concludes that certain ethical judgments in clinical trials, like clinical equipoise, depend upon prior epistemological judgments. If you think that reliable evidence can only come from an RCT, then you are also committed to thinking that clinical equipoise is consistent with a PCT, at least until the standard therapy has been shown superior to placebo. On the other hand, if you accept that reliable medical evidence can come from a variety of other sources besides the RCT, then your judgments about clinical equipoise will tend to be more flexible.

Of course, the further philosophical response to the hard-line epistemic view ought to be: What is the argument? What reason do we have for doubting the reliability of *any and all* evidence gathered outside of an RCT? Worrall's point is well-taken that, yes, there are some examples (or perhaps even many examples) of one-time accepted treatments that were shown to be ineffective through an RCT. But the induction from these particular examples to the universal conclusion that all accepted, non-RCT-tested therapies are suspect is unwarranted.

Moreover, this rather stringent evidential standard is not applied to other sciences, whose ethical stakes are significantly lower. Indeed, it would seem absurd to claim that only controlled laboratory experiments in biology provide real evidence, and that field studies can tell us nothing reliable about the behavior of animals or ecosystems. Griesemer and Wade's (1988) pathways of explanation are an explicit guide to these epistemic relationships between kinds of studies and their justified inferences in biology (cf. §3.3), and it is plausible to think that something similar exists in clinical trials (more on this in §4.4). The fact that clinical trials involve human subjects should cause us to reflect upon the appropriate evidential standards. These may need to be more stringent in some cases, but to reject a huge source of evidence out-of-hand is unnecessary, especially since such a policy would be in direct conflict with the ethical standards already in place within the practice of medicine.

The more widespread defense of evidence-based medicine posits a hierarchy of evidence, with case reports at the bottom and systematic reviews and meta-analyses of RCTs at the top. On this more measured conception, RCTs may provide the best evidence, but there can still be justification for a standard, non-RCT-tested therapy's effectiveness. And as a result, clinical equipoise does not demand an initial PCT.

4.1.3 International ethical regulations

Freedman's notion of clinical equipoise provided a principled, ethical foundation for restricting the use of placebo controls, but this restriction was already represented in a 1964 policy statement drafted by the World Medical Association—the *Declaration of Helsinki*. Written as an international code of ethics to guide experimental research with human subjects, *Helsinki* asserted that "[i]n any medical study, every patient—including those of a control group, if any—should be assured of the best proven diagnostic and therapeutic method" and moreover, any study violating that precept (or any other in the declaration) should not be accepted for publication (World Medical Association 1964).

The *Declaration of Helsinki* has been revised six times since then, and has had an influence on international policy for many years. However, a 2000 revision of the declaration was rejected by both the U.S. Food and Drug Administration (FDA) and Canada's Therapeutic Products Directorate (TPD), the two national drug regulators. The FDA and TPD instead chose to endorse the International Conference on Harmonization's *Guideline on Good Clinical Practice*, whose document E-10 (hereafter "ICH E10"), explicitly permits the use of placebo, even when proven, effective treatment is available (Kimmelman et al. 2009).

The reasons behind this move away from *Helsinki* are in part political (Sampson et al. 2009), but suffice it to say, the overarching thrust of the arguments against *Helsinki*'s restriction of placebo controls, and against clinical equipoise as an ethical requirement, assert that there exists a conflict between the ethical and epistemic needs of clinical research. The chief proponents of this view are Temple and Ellenberg, who claim that placebo controls are not only ethically acceptable, but are epistemically superior to active control treatments. This is due, they write, to every trial's needing assay sensitivity, and an ACET lacking assurances of it.

So influential has their argument been, that a 2008 revision of *Helsinki* now explicitly states that placebo controls are ethically permissible when "for compelling and scientifically sound methodological reasons the use of placebo is necessary to determine the efficacy or safety of an intervention".[3] I will now consider these "compelling and scientifically sound methodological reasons" for the superiority of PCTs.

[3] The complete paragraph from Helsinki 2008 reads: "The benefits, risks, burdens and effectiveness of a new intervention must be tested against those of the best current proven intervention, except in the following circumstances: (1) The use of placebo, or no treatment, is acceptable in studies where no current proven intervention exists; or (2) Where *for compelling and scientifically sound methodological reasons* the use of placebo is necessary to determine the efficacy or safety of an intervention and the patients who receive placebo or no treatment will not be subject to any risk of serious or irreversible harm. Extreme care must be taken to avoid abuse of this option" (World Medical Association 2008, para. 32, emphasis added). Note that condition (1) is entirely consistent with clinical equipoise. Condition (2) is the issue at stake in the next section.

4.2 Sound Methodological Reasoning?

There can be no doubt that if a trial's result is supposed to justify an inference about effectiveness, then it must be the case that the trial had the ability to distinguish between an effective and ineffective treatment.[4] The mere definition of assay sensitivity, as mentioned above, is unlikely to provoke opposition. Nevertheless, as it is described in the ICH E10, assuring assay sensitivity has some peculiar consequences:

> Assay sensitivity is important in any trial but has different implications for trials intended to show differences between treatments (superiority trials) and trials intended to show non-inferiority. If a trial is intended to demonstrate efficacy by showing superiority of a test treatment to control lacks assay sensitivity, it will fail to show that the test treatment is superior and will fail to lead to a conclusion of efficacy. In contrast, if a trial is intended to demonstrate efficacy by showing a test treatment to be non-inferior to an active control, but lacks assay sensitivity, the trial may find an ineffective treatment to be non-inferior and could lead to an erroneous conclusion of efficacy.
>
> When two treatments within a trial are shown to have different efficacy (i.e., when one treatment is superior), that finding itself demonstrates that the trial had assay sensitivity. In contrast, a successful non-inferiority trial (i.e., one that has shown non-inferiority), or an unsuccessful superiority trial, generally does not contain direct evidence of assay sensitivity (ICH E10 2000, pp.7-8).

This passage is unfortunately obfuscatory, especially considering that it is a supposed to be a guideline for clinical research. However, it does clearly suggest that assay sensitivity makes a PCT epistemically superior to an ACET. A successful ACET, meaning that the test treatment performs as well as the control treatment, does not contain "direct evidence" of its own assay sensitivity. Since assay sensitivity is "important in any trial", then this must be bad.

Temple and Ellenberg pick up on exactly this point, writing:

> . . . a study that successfully shows "equivalence"—that is, little difference between a new drug and known active treatment—does not by itself demonstrate that the new treatment is effective. "Equivalence" could mean that the treatments were both effective in the study, but it could also mean that both treatments were ineffective in the study. To conclude from an ACET that a new treatment is effective on the basis of its similarity to the active

[4] Although assay sensitivity is defined in terms of "effectiveness," the ICH E10 is primarily concerned with efficacy. As I discuss below, ignoring the distinction between effectiveness and efficacy glosses over an important difference in the kinds of questions trials seek to answer. I will ultimately argue that just as there are questions about effectiveness and questions about efficacy, there are also two kinds of assay sensitivity: For a more pragmatic trial, it is the ability to distinguish between effective and ineffective treatments; for a more explanatory trial, it is the ability to distinguish between efficacious and inefficacious treatments.

> control, one must make the critical (and untestable within the study) assumption that the active control had an effect in that particular study (Temple and Ellenberg 2000, p.456).

For Temple and Ellenberg, the necessary reliance of an ACET on historical (and hence, untestable) information about the control treatment's effectiveness is problematic. There is always the possibility, they claim, that the active control treatment underperformed in the study, giving the appearance of equivalent effectiveness to the test treatment; leading to a false conclusion. In contrast:

> A well-designed study that shows superiority of a treatment to a control (placebo or active therapy) provides strong evidence of the effectiveness of the new treatment, limited only by the statistical uncertainty of the result. No information external to the trial is needed to support the conclusion of effectiveness (Temple and Ellenberg 2000, p.456).

From these two passages, it might appear as though all Temple and Ellenberg are objecting to is the equivalency or non-inferiority design. Indeed, they explicitly state that superiority to an active control "provides strong evidence of effectiveness". Read in this way, their thesis would only be the limited, and correct, claim above: Observed equivalence in a study is consistent with both treatments being ineffective.

That their claim is not so limited is clear from their case-study. They cite a report of six trials comparing the anti-depressants nomifensine (the test treatment), imipramine (the active control), and placebo. Across all six trials, nomifensine's effects were similar to imipramine. However, when comparing the results of both treatments to placebo, neither significantly outperformed the placebo. Thus, they argue, without a placebo arm in the study, the results would have seemed to erroneously indicate that the test treatment was as effective as the active control.

Temple and Ellenberg take the fact that the control treatment failed to significantly outperform the placebo as evidence of the control's "underperformance" in those studies. That is, they accept that imipramine is, in fact, an effective treatment for depression and would outperform a placebo in other instances. But for whatever reason, it failed to do so in any of these trials. For them, this shows that the trials lacked the ability to distinguish an effective from an ineffective treatment (i.e., imipramine from placebo).

Thus, Temple and Ellenberg's and the ICH E10's methodological ideal is a successful PCT: a trial that shows a new treatment to be superior to placebo. This result establishes the trial's assay sensitivity, and its conclusion of effectiveness is limited only by statistical uncertainty. No external information is needed to interpret its result. An ACET, on the other hand, always leaves open the possibility of its control's underperformance. Any conclusions of effectiveness are suspect for their reliance upon the "external", historical evidence needed to justify that the active control performed as well as it should. Therefore, while we may have ethical reasons for providing all patients participating in a trial with competent medical care, doing good science demands that we give this up.

Or so the argument goes. But before we accept this line of "scientifically sound methodological reasoning", there are four problems with this argument that ought to be addressed:

4.2.1 Efficacy and effectiveness

The first problem is a conflation of efficacy and effectiveness. Both the ICH E10 and Temple and Ellenberg appear to use the terms interchangeably. If clinical trials were only appropriate for answering one kind of question, then this loose use of terms might not matter. However, because some trials are aimed at showing effectiveness and others at showing efficacy, adopting a more pragmatic or explanatory approach, respectively, then the distinction between effectiveness and efficacy is important to maintain.

Insofar as the ICH E10 is taken to apply as an international regulatory standard, its blurring of this distinction is unfortunate and misleading. To be sure, it may be the case that as a drug-regulating document, the ICH E10 is only concerned with issues of efficacy—determining biological action and safety. Yet, it is not the case that all trials falling under its purview are designed to answer efficacy questions. As discussed in §4.1, a more pragmatic trial, which seeks to answer a question about comparative effectiveness, actually has different epistemic relationships from one which seeks to answer a question about efficacy. A PCT may be more useful for isolating the biological effect of a new treatment, but so long as a placebo is inconsistent with competent medical care, it makes the trial less like clinical practice, and thereby weakens its external validity. Specifically, a PCT does not answer the physician's question: "How does this new intervention compare to what we are already using?"

It is also important to keep in mind that there is still an important role for pragmatic trials in a regulatory context, and collapsing the distinction between effectiveness and efficacy risks ignoring a critical methodological consideration about the question a clinical trial is supposed to answer. Furthermore, in cases of serious illness, for which an effective standard therapy is available, a PCT is simply not an option. A pragmatic trial is thus necessary to determine either the effectiveness or the efficacy of any new treatment. Therefore, even in a regulatory context, the concept of assay sensitivity needs to be broad, and yet precise, enough to cover both kinds of trial questions.

To further reinforce the point, in the face of exploding health care costs in the U.S., the American College of Physicians has recently been calling for more "comparative effectiveness" studies. These are studies which evaluate the relative clinical *effectiveness*, safety, and cost of two or more medical interventions. Kirschner et al. (2008) have even gone so far as to advocate for a national comparative effectiveness program, arguing that the lack of such studies "interferes with the ability of physicians and their patients to make effective, informed treatment choices that meet the unique needs and preferences of the patient and facilitate the ability of payers to optimize the value of their health care expenditures" (Kirschner et al. 2008, p.956). As federal health coverage programs expand in the U.S. (and elsewhere), the need to manage health care costs only becomes more important. If comparative effectiveness programs are successful at managing such costs, then this increases the need for pragmatic trials, and places even more weight on our claim that federal regulators ought to expand their guidelines to better encompass pragmatic trials.

4.2.2 Epistemic and ethical problems with placebos

The second problem concerns the science of placebos themselves. As Freedman (1990), and more recently, Howick (2009), have shown, many of the supposed epistemic advantages of using a placebo control are illusory. The so-called "placebo effect" is not in fact one, consistent effect, but is a wide-ranging, variable effect that differs, not only from one trial to the next, but also depending upon whether the placebo is an injection, capsule, red tablet, blue tablet, etc. (Shapiro 1970, Spiro 1986). The idea that a placebo provides a consistent baseline measure by which to determine the test treatment's absolute efficacy is simply false. So too is the idea that a treatment's biological effect can be accurately measured in a PCT by subtracting from its observed effect the effect observed in the placebo arm. This "assumption of additivity," as Howick (2009) shows in some detail, does not generally hold, and thus, PCTs, just like ACETs, can provide, at best, only relative measures of treatment efficacy or effectiveness.

Temple and Ellenberg's worry about an active control's underperformance, and the resulting epistemic uncertainty, also applies to PCTs. In a randomized, double-blind PCT, neither the physician nor the patient is supposed to know whether or not the patient is receiving the test treatment or the control. By equalizing patient expectations of improvement between the two arms in this way, any difference in outcomes in the test treatment arm can be reasonably attributed to the biological effects of the treatment. However, if patients become unblinded, a differential expectation between the two arms of the studies is introduced. The patients in the placebo arm become aware of the fact that they are receiving a placebo rather than an active treatment, and their expectation of improvement is undermined.

Underperformance of the control due to unblinding is therefore a potential problem for all clinical trials. But it is important to note that unblinding does not pose the same problems for ACETs as it does for PCTs. In an ACET, the knowledge of which treatment a patient is receiving may introduce certain observational or reporting biases, but it does not introduce a qualitative difference in expectation. An unblinded ACET can indeed suffer from a differential expectation between the two treatment arms, since patients may be biased by believing either the older or newer treatment is more effective, but this is quite different from the qualitative distinction introduced into an unblinded PCT, in which the patients in the control arm learn that they are receiving nothing. Thus, the critical distinction we make here is between a *quantitative* differential in patient expectation as a result of learning that they are receiving the standard, effective treatment in an ACET versus a *qualitative* difference in patient expectation as a result of learning that they are receiving nothing in a PCT. To be sure, both situations as a result of unblinding are to be avoided in a well-designed and well-executed trial; however, given the real dangers of unblinding, the distinction between the two is important to recognize.

Speaking of the real dangers: An analysis of PCT results published between January 1998 and October 2001, found that in trials for general medicine, only 7 of the 97 sampled trials assessed the success of blinding, and in psychiatric trials, only 8 of 94. Of the 15 total trials that did report, 9 of them found that blinding was imperfect (Fergusson et al. 2009). The authors rightly conclude:

> Our examination of the success of blinding challenges the notion that placebo controlled trials inherently possess assay sensitivity. Clearly, there is a failure among investigators and journals in reporting the success of blinding. . . . This deficiency in reporting translates in a paucity of evidence that a placebo ensures a “clean” control (Ibid., p.434).

Given then that patient expectation of improvement is undermined by learning that they are receiving a placebo, it is critically important for a PCT to minimize the possibility of unblinding. This can be done by using placebos that reproduce some or all of the side-effects of the test treatment. However, as Howick points out, the patients in such a placebo arm are not only being deprived of competent medical care, they are now actually being harmed by their participation in the study (Howick 2009, pp.36-37).

Concerns about the acceptable harms to patients in the placebo arm of a clinical trial is a long and on-going debate in the research ethics literature. Freedman et al. (1996) argue that few dispute the unethical use of placebos in cases where irreversible harm or mortality are the dangers, rather, the debate is really about what constitutes “reversible harm” and “serious morbidity”. They discuss the example of a PCT for the treatment of chronic schizophrenia in which 66 percent of patients in the placebo group relapsed during the 9 month trial, versus only 8 percent in the treatment group (Curson et al. 1986). They write:

> Does a drug company, regulator, investigator, or institutional review board (IRB) have the right to judge that the direct harm suffered by the placebo group during the

> course of the trial period is not serious? In some cases, the point should be clear: many sufferers of schizophrenia or depression consider their malady a fate worse than death. . . . The question becomes one of defining "seriousness" and "severity". Discounting the psychological pain, the disruption of relationships, and the heavy burden on families during a period of nontreatment as not being serious or severe enough to mandate treatment demonstrates an unacceptable disregard for the well-being of psychiatric patients and those responsible for their case (Freedman et al. 1996, pp.253-254).

The point here is not that placebos are epistemically useless or always unethical. Rather, the point is that they are not epistemically or ethically unproblematic, and the discussion in Temple and Ellenberg and the ICH E10 does not provide an adequate treatment of these issues before concluding that PCTs are methodologically superior to ACETs. The duty of care cannot be shrugged off so easily.

4.2.3 Duhemian underdetermination

The third problem with the assay sensitivity argument takes us back to where this entire thesis began: Duhemian underdetermination. James Anderson (2006), in his critique of Temple and Ellenberg's argument, points out how their concern that an ACET relies upon information external to the trial is just a species of underdetermination. As he rightly observes, a clinical trial is a kind of experiment, and like all experiments, a number of assumptions are needed for the interpretation of its results. Contrary to Temple and Ellenberg's contention, the *majority* of the assumptions involved in any experiment are external, in the sense of not being directly tested in the trial itself. A PCT is no different from an ACET in this regard. Both kinds of trials require external assumptions or information.

Indeed, apart from underperformance due to unblinding, as discussed above, Temple and Ellenberg's general worry about possible treatment underperformance can amount to nothing more than an underdetermination claim. If a trial's result fails to show a clinically or statistically significant difference between the test treatment and active control, it could be, as Temple and Ellenberg fear, because the active control treatment underperformed in that trial. It could also be because the two treatments are not appreciably different. Or it could be the result of some observational bias or measurement error on the part of the physician-researchers. A well-designed study surely aims to control for possible confounding and misleading factors, and the working assumption is that none of them will be important enough, if the trial is properly conducted, to bias the final result. Nevertheless, a judgment must still be made by the scientist. She will always be called upon to use good sense in interpreting the evidence, no matter how dramatic her result.

For Temple and Ellenberg's argument to run, it would need to be the case that these interpretive judgments could never be adequately justified. But such a radical skepticism requires another argument than the one they provide, perhaps one more akin to Worrall's hard-line empiricist. The basic observation that an experiment relies upon "external" information for its interpretation only illuminates the need, as Duhem observes, to exercise good sense.

4.2.4 Two epistemic contexts

Finally, the ICH E10 explicitly, and incorrectly, describes assay sensitivity as a property of a trial to be discerned in light of its result. Assay sensitivity is a prospective determination about what a trial can show. In the more pragmatic case, it is a trial's ability to distinguish an effective from an ineffective treatment. In the more explanatory case, it is a trial's ability to distinguish an efficacious from an inefficacious treatment. Therefore, the claim that the implications of assay sensitivity change depending upon the intent *and* result of the trial is false. If assay sensitivity is a trial's "ability to distinguish", then this is discernible before the trial has been executed.

The idea, made explicit in the ICH E10 and implicit in Temple and Ellenberg, that "when two treatments within a trial are shown to have different efficacy (i.e., when one treatment is superior), that finding itself demonstrates that the trial had assay sensitivity" is similarly incorrect. If two treatment arms within a trial are observed to have different efficacy, this may indeed provide some retrospective evidence that the trial was assay sensitive, but this does not rule out the possibility of a false difference, i.e., a type-I error. A poorly designed or executed trial may give the appearance of different effects when in fact the two treatments are not appreciably different. If assay sensitivity were nothing more than a trial whose result merely *appears* to suggest that two treatments have different efficacy or effectiveness, regardless of the trial's quality, design, or internal conduct, then it would be completely trivial and surely not the critical epistemic property it is taken to be.

To clarify this issue, I propose a distinction between two epistemic contexts:

1. Trial-as-designed: What is an experiment *capable* of demonstrating?

2. Trial-as-executed: What does an experiment, *in fact*, demonstrate?

One ought to be able to answer question 1 before knowing the experimental result. What an experiment can demonstrate in principle is addressed by philosophers and methodologists of science. The discussion of Duhem and underdetermination has already touched on this point. But it bears repeating that what a particular experiment, like a clinical trial, is *capable* of demonstrating can be assessed in advance of carrying out the trial. Factors like adopting the appropriate orientation (i.e., pragmatic or explanatory), asking the relevant question (i.e., effectiveness or efficacy), making the design assumptions explicit, and minimizing bias are all relevant to this judgment.

Question 2, on the other hand, can only be assessed after an experiment has been completed and the result is known. It is answered in the course of the interpretation and discussion steps. While the issues of design and interpretation are related (in ways to be illuminated below), they are also separable. In the case of a clinical trial, whether or not its result justifies the inference that a difference exists between two treatments depends upon more than just the findings. It also depends upon the statistical uncertainty, the success of blinding, the internal conduct of the trial, and all of the other auxiliary assumptions involved in the design and execution of the experiment. It may be that the trial's result, in light of other assumptions, is less decisive than hoped, but this does not imply that the study as a whole *lacked the ability* to be decisive.

Thus, we can say that so long as a trial is judged to be well-designed, the problematic scenario of a treatment's underperformance in a trial is not a problem for the trial-as-designed, but a problem for the trial-as-executed. Temple and Ellenberg's skepticism is thus not even relevant for determining assay sensitivity—properly understood to be a prospective judgment about the trial-as-designed. After a trial is executed, it may be that the results radically underdetermine the appropriate inference. But whatever the cause or causes of these difficulties, they are due to the study's sample or conduct, not its design.

4.3 Assay Sensitivity Revised

Having now rejected Temple and Ellenberg's argument for the methodological superiority of PCTs, we can return to questions about assay sensitivity itself. Adopting the straightforward understanding of assay sensitivity, we can say that a pragmatic trial has the ability to distinguish between effective and ineffective treatments, or an explanatory trial has the ability to distinguish between efficacious or inefficacious treatments, when it is well-designed. What then is meant by a "well-designed trial"? Aside from conforming to the methodological standards like random assignment to treatment groups, double-blinding, and so on, a well-designed trial, just like any well-designed experiment, is one whose auxiliary assumptions and potential biases are recognized. As Duhem points out, an experiment's result is potentially limited by all of the auxiliary assumptions used in its experimental design, and as such, when the scientist calls upon her good sense to interpret that result, the more she understands and recognizes the limitations of the untested assumptions involved in her experiment, the better (Duhem 1906, p.216-218).

Drawing an analogy between trials and scientific models is helpful here. Models, like experiments, are an important part of the scientific process, useful for both predictions and explanations. Yet, they are always false in one way or another. Models introduce simplifications and idealizations, ignoring or distorting some features of the systems under study in order to better capture the features of interest. But because these simplifications make the models false, at times they will fail to correctly capture the behavior of the system.

Happily, even a model's failures can still be quite useful. Wimsatt offers a prescription on this very point:

> The primary virtue a model must have if we are to learn from its failures is that it, and the experimental and heuristic tools we have available for analyzing it, are structured in such a way that we can localize its errors and attribute them to some parts, aspects, assumptions, or subcomponents of the model (Wimsatt 2007, p.103).

As is now familiar, "localizing" errors in a model means understanding the heuristics used in its construction and knowing how to recognize when those heuristics have led us astray. Underdetermination is still a factor, since models do indeed confront evidence as a whole. Nevertheless, Wimsatt's point is that well-designed models can still be broken down into their constitutive heuristics in order to have their errors diagnosed and corrected.

For example, take Wimsatt's "control" heuristic of experimental design:

> ζ_1 : When designing an experiment, the environment (or environmental variables) should be kept (or assumed to be) constant (Wimsatt 2007, p.83).

Applying ζ_1 is a good way to test for relationships between variables in a system, but it will ignore (and hence bias against observing) any relationships between system variables and extra-systemic factors. All experiments apply the control heuristic to some extent, but the degree to which the experiment is controlled can have an important effect upon the result.

Fuks et al. (1998) provide an illustration of this point in clinical trials. They find that eligibility criteria (e.g., age restrictions or circumstances of prior treatment), set during the trial's design, are the largest factor in explaining low patient recruitment rates for clinical trials, frequently excluding over 40% of the potential patient population (Fuks et al. 1998, p.69). They compare and contrast changes in eligibility criteria between two clinical research programs over a 20 year period: The National Surgical Adjuvant Breast and Bowel Program (NSABP) and the Pediatric Oncology Group (POG). Between 1972 and 1992, the NSABP conducted 22 breast cancer trials. During that time the number of eligibility criteria increased from 21 to 44. This is in contrast to the seven POG trials conducted over the same period, for which the number of criteria went from 6 to 12. Not only did the NSABP have more criteria to begin with, it added criteria at a greater rate, and once added, a criterion was very unlikely to be removed later (Ibid., pp.72-73).

Fuks et al. characterize the difference in the approach to eligibility criteria as between an explanatory (NSABP) and pragmatic (POG) research program, but we can also characterize it as differing in the degree to which ζ_1 is applied. Trials which control for more variables (e.g., excluding patients with possible confounding conditions who would nevertheless be treated in practice) are more explanatory. They more fully exclude possible confounding influences, and therefore, better support inferences of efficacy. Conversely, trials with minimal entry requirements, controlling for fewer circumstantial confounders, are more pragmatic. They more closely mirror conditions in clinical practice and can better support inferences to effectiveness.

The point is not to avoid ζ_1 altogether in designing an experiment, since it is clearly important for answering explanatory questions. Rather, the point is to be aware of it as a heuristic and understand the biases it introduces. In Fuks et al.'s analysis, the NSABP was too controlled; too explanatory given that the aim of the program was to inform breast cancer treatment practices. The POG program, in contrast, ran trials of equally high quality, but because of their less restrictive eligibility requirements (i.e., more pragmatic approach), were able to recruit more patients and better inform clinical practice. The lesson here can be summarized with the meta-heuristic, Z_1:

Z_1 : For pragmatic trials, limit the use of ζ_1.

What makes a well-designed model well-designed is that it is possible to learn from its failures. When it breaks down, if we can still determine which of its simplifications is causing the problem, then we have learned something about its limitations, and are in a better position to draw well-supported inferences from the model's result. This is made possible with a working understanding of the heuristics involved in the mode's construction. Extending this line to clinical trials, we can say that a well-designed trial, and therefore also an assay sensitive trial, is one whose potential biases and errors can be properly attributed to the heuristics or simplifications of its design. Fuks et al.'s analysis shows how this can be done retrospectively.[5] The PRECIS tool, introduced in §2.2.1, shows how this can be done prospectively. A researcher can refer to table 2.1 to assess where the trial's assumptions lie on the pragmatic-explanatory continuum, and judge whether this is appropriate given the question they want to investigate.

In addition to the table, a PRECIS graph, like that shown in figure 4.1, provides a visual summary of a trial's orientation. The example plot provided here represents a mostly "mixed" trial, partly pragmatic and partly explanatory.

[5] This speaks to an earlier point that although the contexts of a trial-as-designed and a trial-as-executed are distinct, they are still related. Understanding how a trial or series of trials was designed can help to inform judgments about the quality of evidence it provides.

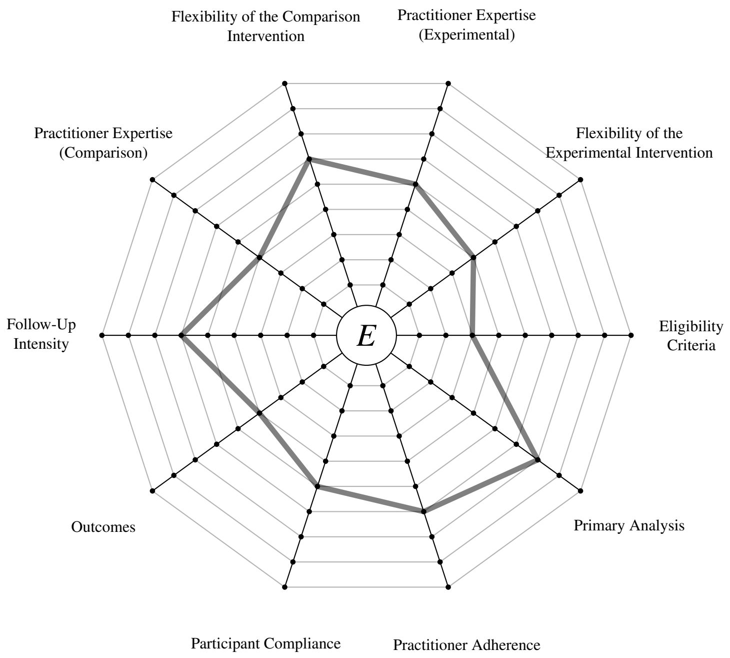

Figure 4.1: PRECIS graph of a mixed trial

For the trial-as-designed context, the PRECIS graph helps to inform two questions: What kind of question is the trial designed to answer? How does this match up with the stated aims? Given a particular plot on the graph, we can judge the kind of question it will be well-suited to answer. For example, the "mixed" plot in figure 1 might correspond to an early effectiveness study. It adopts some explanatory assumptions, but is pragmatic enough to be informative for clinical practice.

But there are other, yet more basic questions about whether or not a trial has covered all of its methodological bases, so to speak. The CONsolidated Standards Of Reporting Trials ("CONSORT statement") is a tool designed explicitly for this purpose (Begg et al. 1996, Moher et al. 2001). Built upon a set of methodological standards, the CONSORT statement is a checklist and flow diagram addressing information commonly associated with biased estimates in treatment effect, reliability, and relevance of findings. In recommending trialists use the checklist, CONSORT aims to assist the medical community in interpreting and judging the quality of published results. Like PRECIS, it is intended as a fluid document, to be revised on an on-going basis as medical experts and biomedical publications evaluate the merits of its criteria and content (Moher et al. 2001).

Tools like CONSORT and PRECIS illuminate the fact that although assay sensitivity, in the sense of good design, is a judgment about trials-as-designed, it is still a relevant experimental feature when it comes to judging the trial-as-executed. By making explicit the heuristics and assumptions employed in the design, the clinical researchers not only improve their design, they also elucidate their interpretation, and provide opportunities for alternative judgments and better informed discussion of results within the expert community.[6]

Indeed, CONSORT was intended as a trial quality reporting tool (i.e., to assess the trial-as-executed), but it has begun to be used as a guide to trial design. The point I wish to emphasize here is that thinking about assay sensitivity as meaning a well-designed trial, made explicit with heuristic tools like PRECIS and CONSORT, helps to distinguish between the trial-as-designed and trial-as-executed contexts, while simultaneously illuminating an important relationship between the two. A well-designed study makes its assumptions and potential biases explicit, and in so doing, is better equipped to protect against faulty interpretations. Indeed, faulty interpretation was the legitimate concern behind Temple and Ellenberg's assay sensitivity argument. As I have now shown, the way to address it is not to emphasize placebo controls, but to delve more deeply into the methodological heuristics of clinical trials.

4.4 Robustness and a Clinical Research Program

All of the above goes to show how a well-designed experiment, in making its assumptions explicit, improves the possibilities for justified interpretations of its evidence. This is a general feature of scientific experiments and has nothing to do with what kind of experiment it is. In clinical trials, this means that PCTs and ACETs are, *in principle*, on equal epistemic footing. So long as they are well-designed, the evidence obtained from each can legitimately justify inferences.

Just what kind of inference is justified by a trial depends, in part, on the question it seeks to address. But it is not enough to conclude, on the basis of a single trial, no matter how well-designed, that a particular treatment is either effective or efficacious. Analysis of the trial-as-executed context must therefore look beyond the single trial to judge evidential contribution in light of other studies.

[6] One important characteristic in both the design and reporting of any trial is its *power*. When clinical trials are designed, they must determine how many subjects to enroll. This is done by estimating the difference in effect size (i.e., the expected difference in outcomes between the two treatment arms) and choosing a type-I (false positive) and type-II (false negative) error rate, represented by the variables α and β, respectively. α is usually set by convention to 0.05 or 0.01. $1-\beta$ is the power of a study, and is usually set by convention to 0.80 or 0.90. Part of a well-designed study is getting this initial effect estimate correct. Contra Temple and Ellenberg, this requires historical information about the control treatment's effect size.

The post-hoc power of a study is likely to change from this initial value of 0.80 or 0.90, as the difference in effect size varies from the initial estimate or the final enrollment differs from the calculated number. On this point, it is worth recalling the study of six antidepressant trials cited by Temple and Ellenberg. An earlier presentation of this same data from Temple (1983) includes the post-hoc power of each study to detect the observed difference, which ranged from a mere 0.09 to a maximum of 0.40. Considering the difference between those and the conventional standards, it would have been clear to anyone interpreting the date that all six of those studies were under-powered, and any conclusions drawn from them poorly supported.

Expanding on the above comparison between models and trials, just as a single scientific model is insufficient for establishing a robust result, since there is no guarantee that its result is not an artifact of one of its assumptions, so too is a single trial inferentially limited. As I argued in chapter 2, a result is only robust when it is invariant across a variety of experimental or modeling heuristics. As I have already noted, experiments, just like models, employ heuristics and thereby introduce certain biases. Controlling for these biases in a single experiment is essential to good design and valid interpretation, but there is only so much that one experiment can control. A robust pattern, invariant across multiple experiments, offers another way to control for and detect biases. Since different experimental heuristics introduce different biases, a result which is invariant across a series of experiments, all varying in their design heuristics and potential biases, is more robust, and hence, more reliable. This is in contrast not only to the result of a single experiment, but also to a result that is merely invariant across a series of identical experiments. In the latter case, it is still possible that the result's invariance is due to the shared biases of the common design. In yet another reversal of Temple and Ellenberg, the less self-contained a trial is, the more it incorporates or corroborates previous work, the more robust and less limited are its conclusions.

As I also argued in chapter 2, robustness figures into our experimental epistemology in two ways: First, it is a property of results. Results shared across different investigations, each utilizing different heuristics are less likely to be artifacts, and are therefore more reliable. Second, robustness is an ideal aim of research. Establishing a robust property is not a fool-proof affair. Culp (1994) and Wimsatt (2007) both provide examples from the history of biology where judgments about a result's robustness were later shown to be incorrect. These examples, much like the historically ineffective medical interventions mentioned by Worrall (2008), speak to some of the difficulties in testing for robustness. It can seem as though we have sufficient evidence to justify a robustness claim, but the possibility of a hidden, shared bias always remains.

This ever-present possibility of pseudo-robustness explains why robustness is an *ideal* aim of research. We cannot eliminate doubt or error, but we can still work to minimize it. To do this, we design experiments so that they optimally contribute to our understanding, perturbing heuristics and design assumptions in an effort to "weed out" bias. Genuine robustness can therefore be thought of as the ideal result from a thorough and systematic exploration of problem space.

The "shape" or dimensions of the relevant problem space, as well as, the qualifiers "thorough" and "systematic", will all need to spelled out in more detail for any specific domain. Figure 4.2 provides a model of how this might look for a medical research program. The long experimental chain from a new intervention's initial laboratory testing to its approval for use

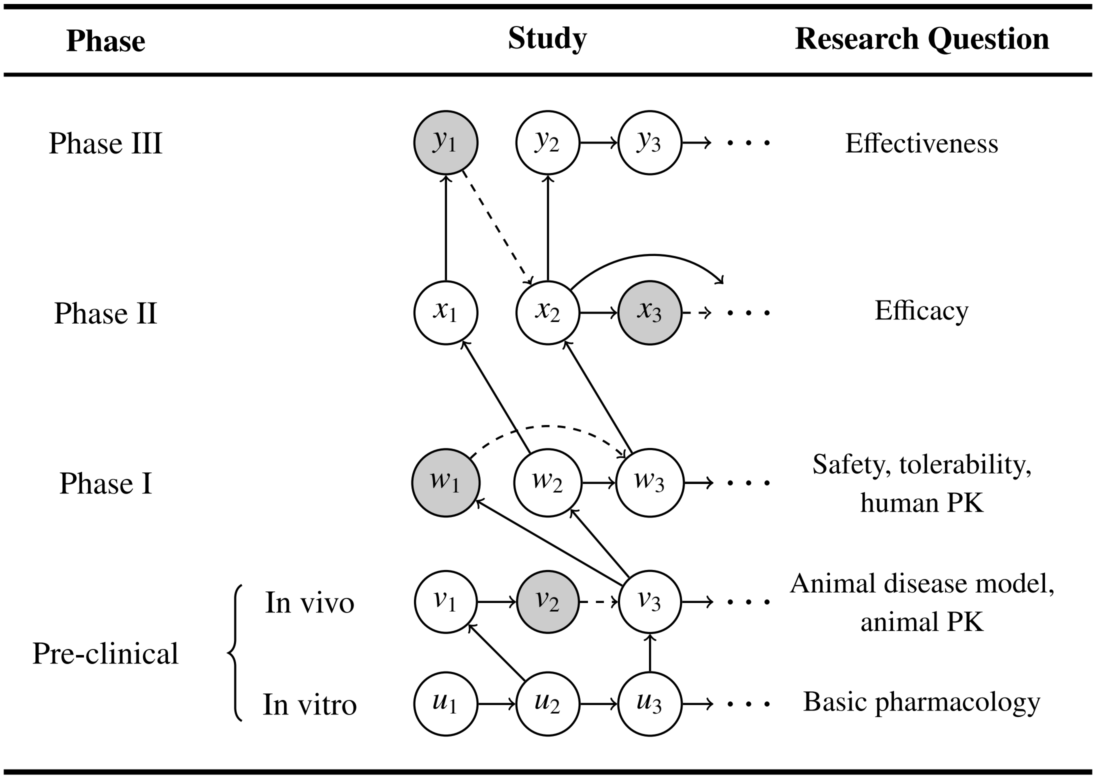

Figure 4.2: Model of problem space for a clinical research program

in clinical practice is a complex affair, requiring several different kinds of experiments and hypotheses, as well as, layers of epistemic and ethical considerations once human research subjects are involved.

Although the standard "phase" divisions (i.e., pre-clinical, phases I, II, III) are in some cases arbitrary, we can use them here to make some robustness considerations explicit. Research into a new treatment is not multiple tests of a single hypothesis. Rather, it is a series of different tests and hypotheses, some designed to answer explanatory questions and others to answer pragmatic questions: The pre-clinical and in vitro studies answer questions about basic pharmacology; the in vivo studies answer questions about animal disease models and pharmacokinetics (PK). Phase I studies are designed to answer questions about safety, tolerability, and human PK. Once a dosage is found that strikes a desirable balance between effect and toxicity, it is put to use in a larger study population for the phase II study. The second phase is directed toward a more rigorous evaluation of the treatment's biological activity, seeking to show the magnitude of the treatment's effect and the number and kind of adverse events. The overarching aim of the phase II study is to inform a decision about whether or not the experimental treatment is appropriate for a large, practice-changing phase III study.

The ultimate aim of the whole research program is, of course, to establish an intervention's effectiveness. But along the way, there are questions about efficacy, safety, and pharmacology; and each of these prior questions, in addition to the effectiveness question, demand some degree of robust evidence. There are, therefore, two dimensions of robustness: (i) *consistency*, or what we might think of as "horizontal robustness", which is invariance of results at the same stage of research (e.g., invariance across all in vitro studies); and (ii) *congruence*, or "vertical robustness", which is invariance of results at different stages of research (e.g., invariance across in vitro, in vivo, and into later phases). Both of these kinds of robustness are sought throughout the research process. Indeed, inconsistency within or incongruence between the phases can be reason enough to discontinue research with a particular treatment (e.g., because it appears unpromising in humans), to re-evaluate experimental designs (e.g., due to suspicions of bias), or to perform additional tests (e.g., in an effort to resolve evidential discordance).

The network of nodes in figure 4.2 helps to make these two dimensions of robustness explicit. The white nodes correspond to studies with positive results; the shaded nodes to studies with null results. The arrows between studies represent temporal and epistemic relationships. In the constructed example, we can say that the phase I study, w_2, relied upon the earlier, animal model results from v_3, which in turn relied upon the earlier null study, v_2, which relied upon yet earlier result from v_1, and so on. In this way, the paths and connections of the network not only represent the actual medical practice of later studies citing and incorporating prior work, but it also tracks a history of evidence and justification across the research program.

Consistency is represented in the model when all the studies (or some sufficient number of them) at a single phase of research show positive results. In the example, the in vitro studies (u_1, u_2, u_3) were all positive. This would be evidence that we have a robust understanding of the intervention's basic pharmacology, and, we might think, adequate justification for moving to in vivo studies (although more on this question below). At the in vivo phase, the first study, v_1, was positive while the second, v_2, was not. We use the dashed arrow, as between v_2 and v_3, to represent a difference in epistemic judgment vis-à-vis robustness. Where a solid arrow can be thought of as direct evidence of robustness, the dashed arrow represents evidence of the limits on, or conditions of, that robustness. The dashed arrow invites further questioning about why the inconsistency (or incongruence) might have been observed, and may suggest that questions earlier in the research path need to be re-evaluated. This kind of experimental refinement is captured in our example by having v_3 informed by both a later in vitro study, u_3, and the null study, v_2.

Congruence is represented in the model by the "vertical" path of research. All of the positive studies, following the path from u_1 to v_3 to w_3 to x_2 to y_3, tell a story of congruence; an overall understanding of the treatment's pharmacology and a demonstrated pattern of efficacy, safety, and effectiveness.

While there is certainly much more information that could be built into the model, even this limited representation allows us to draw some interesting conclusions. To begin with, we can ask questions about the connectedness of the network. This translates into the question: How much do later studies in the research program draw on the work of previous studies? Or conversely: How much are studies used to inform later research? In our example, there is no isolated study that does not make use of previous work (excepting the very first in vitro study, u_1). There is also no study that does not inform later work. Although we have limited our example to three studies at each phase and generally avoided having two studies of the same phase going on at the same time (w_1 and w_2 are the exception to this), these are most certainly idealizations for the purposes of introducing the model. There is no reason to suppose that an actual research program would adhere to either limitation.

Yet, it is worth reflecting on the methodological rules to which such a grand research program *should* adhere. A hard limitation on the number of studies at any particular phase seems obviously imprudent, but there might be good reasons to limit the number of concurrent studies taking place, or the number of positive studies executed at one phase before moving on to the next phase. In the example, v_3 is the second positive, in vivo study, providing the justification for the concurrent phase I studies, w_1 and w_2. The null result in w_1 is still useful for designing w_3, but since the cost of each study increases substantially as a program moves "up" through the phases, one could plausibly argue that w_2 would have been improved had it been performed after w_1 (when it could also have drawn on the results of more phase I studies).

But one could also run an argument the other way: If time, rather than cost, is the more pressing factor, then there may be good reasons to run concurrent studies, hoping that for each null study, something valuable can still be used to improve future research. The path from w_2 to x_1 to y_1 to x_2 reflects this second methodology. The phase II study, x_1 began after only one successful phase I study (w_2), and in turn, was the sole phase II justification for the phase III trial, y_1. Although y_1 provided only a null result, assuming that it was assay sensitive, it *could* have been positive. Nevertheless, its result may still have provided methodologically useful information (e.g., about issues of dose response or treatment sub-population) that could be incorporated into the design of the subsequent phase II study, x_2.

The point here is not to provide all of the possible prescriptions and judgments needed to fully exploit the model. It is rather to suggest how one might structure an analysis of the problem space for medical research. Throughout this chapter, I have essentially been focusing the discussion on phase III trials. Insofar as they seek to inform and change clinical practice, I have argued that they should be asking pragmatic questions about the experimental treatment's effectiveness. But to justify the pragmatic question, there will need to be evidence from the earlier phases that the experimental treatment is safe and efficacious. In other words, the ev-

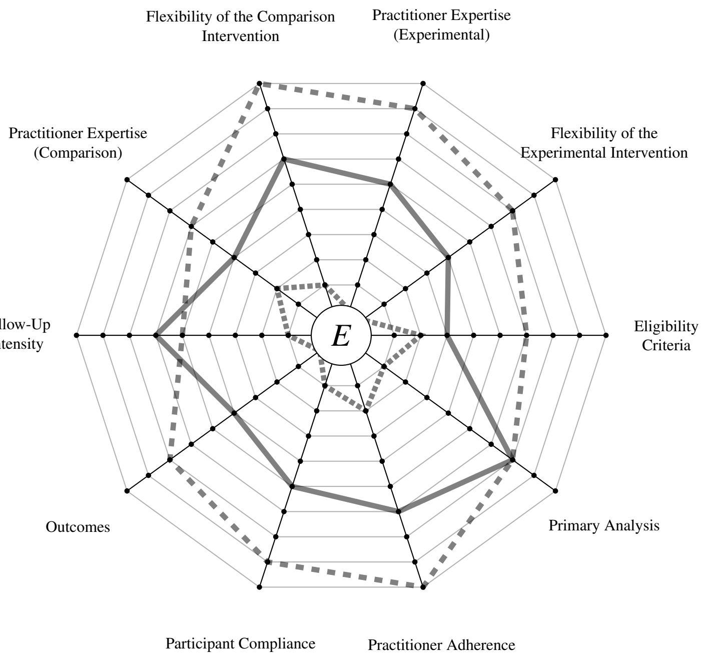

Figure 4.3: Robustness PRECIS graph for 3 Trials

idence that a state of clinical equipoise ought to exist—ethically necessary to justify a phase III trial—relies upon these early studies. A robust clinical result will, therefore, be one that is congruent across all phases of research. While regulators have been most interested in explanatory questions, our model of the problem space makes explicit that the ultimate end is actually a different question—a pragmatic question about the effectiveness of new treatments. Information about a treatment's efficacy is certainly important, but insofar as it does not readily inform decision making in clinical practice, then there is a strong case to be made that regulators ought to shift their focus to be more in line with the needs of physicians. Indeed, the push toward comparative effectiveness studies is a step in exactly this direction. In the context of our model, this is a demand for increased evidence of consistency at the phase III level.

On that point, we can refer back to the PRECIS graph's ability to represent judgments of consistency at the phase III (or even the late phase II) stage. Figure 4.3, for example, illustrates a comparison of three trial designs. In a conceptual sense, we can imagine each node in the problem space from figure 4.2 as corresponding to a particular PRECIS graph (i.e., you can imagine "zooming" in on the node until the additional information about its design dimensions can be "seen"). By "overlaying" the plot from each trial onto a single graph, we can get a picture of how well-explored is this "lower-level" space of trial design. There are thus multiple levels of analysis at work: The higher-level analysis of the overall research problem space—represented in figure 4.2 by the network of studies within and between phases; and the lower-level analysis of trial design space—represented for later phase trials by the PRECIS graph.

The dotted line in figure 4.3 thus corresponds to a late phase II or early phase III explanatory trial.[7] It will most reliably tell us about the treatment's efficacy, but may also be suggestive of effectiveness. The solid line represents a "mixed" trial, explanatory in some respects, pragmatic in others, providing more evidence of effectiveness. Finally, the dashed line around the edge represents a largely pragmatic trial, designed to give us the best evidence of effectiveness.

The ideal is to have a positive pattern of treatment effects and performance, consistent across all three of these studies. By varying the degree of control, shifting from explanatory to pragmatic assumptions, we gradually weaken the likelihood that the positive result is due to some artifact of the investigation. As a result, we can be confident that the treatment is truly effective.

Thinking about the problem space of possible research questions and trial designs with these conceptual tools makes it seem almost absurd to conclude after a single trial that a treatment was effective (or not). On the other side of the coin, a robustness PRECIS graph, nearly filled with trials of different kinds, perturbing their assumptions along each dimension, provides a vivid illustration of a thoroughly explored research question. Combined with the network model from figure 2, we can be explicit about what a "thorough" and "systematic" exploration of problem space means for medical research: Each study needs to build upon previous work; perturbing its assumptions to explore the different dimension of experimental design appropriate to its phase.

Chapter Summary

While regulators may be most interested in explanatory questions, I have argued here that there are other kinds of questions that clinical trials are capable of answering—pragmatic questions about the effectiveness of new treatments. These are questions for which the use of placebo controls are not well-suited, yet they are exactly the kinds of questions that physicians care about the most. Since information about a treatment's efficacy does not readily inform decision making in clinical practice, there is a strong case to be made that regulators ought to shift their focus to be more in line with the needs of physicians. Indeed, the push toward comparative effectiveness studies is a step in exactly this direction. Yet, the issues here are quite contentious. The role that regulatory agencies ought to play in prescribing the methodology for testing new

[7] The methodological distinction between later phase II trials and early phase III trials is a blurry one. For some interventions there may be no appreciable difference here at all, except for the difference in research question. Therefore, it may be entirely appropriate to represent phase II and III trials along the same pragmatic-explanatory dimensions.

treatments is not obvious. Although this is an important question, it is beyond the scope of this thesis.

Instead, the aim has been to extend concepts from the philosophy of science and develop a framework for understanding the epistemology of robustness in clinical trials. I have argued that the understanding of assay sensitivity presented by Temple and Ellenberg, codified by the ICH E10, and adopted by the FDA and TPD, collapses the distinction between a treatment's effectiveness and efficacy, as well as, the epistemic considerations relevant to trials-as-designed versus trials-as-executed. The principles of "sound scientific methodology," described in the *Declaration of Helsinki*, do not rule out the possibility of using an active control treatment in clinical research; and therefore, there is no need to abandon our ethical obligation to provide all patients participating in a clinical trial with competent medical care. It is possible to satisfy both the epistemic norms of science and the ethics of care simultaneously.

In attempting to remedy this state of confusion, I have argued for a straightforward understanding of "assay sensitivity" as a prospective judgment about a trial's design. To revise the guidelines from the ICH E10: If a pragmatic trial, designed to answer a question about effectiveness, is assay sensitive, it will be capable of distinguishing between an effective and ineffective treatment. Similarly, if an explanatory trial, aimed at answering a question about efficacy, is assay sensitive, it will be capable of distinguishing between an efficacious or inefficacious treatment. In both cases, judging a trial to be assay sensitive requires that it is adequately powered, makes its design assumptions explicit, and aims to minimize bias through its contribution toward a consistently and congruently robust result. Thinking about assay sensitivity in this way has epistemic advantages for both the trial's design and its interpretation, not the least of which is that it becomes much easier for review boards and others in the expert medical community to recognize the influence of known biases or problems, and diagnose them as stemming from the trial-as-designed or the trial-as-executed.

As I have shown, the worries about assay sensitivity in trials are really just a species of worries about how to design models and experiments and how their design bears on interpreting their results. No experiment is an atomistic test, capable of standing on its own. A clinical research program is a complex web of hypotheses, tests, and judgments, both epistemic and ethical. The concepts of experimental heuristics, their relationship to bias, and the need for robust results are all important contributions from the philosophy of science that offer a novel and powerful way to think about the epistemic aims and structure of clinical research.

Chapter 5

The Aim and Structure of Clinical Research

I ended the last chapter by concluding that no clinical trial is fully self-contained. The more its interpretation can draw on results from other well-designed studies to support its conclusion, the better. This speaks to the importance of robustness in clinical research, both as a property of shared results and as a research meta-heuristic, i.e., a process of perturbing experimental heuristics in order to maximize reliability. Also in the last chapter, I suggested a model for thinking about robustness in a clinical research program: a network of nodes and edges, representing trials and their epistemic relationships, respectively. The aim of the model is to provide a visual representation of how a research program has explored its problem space, potentially sharpening questions about how connected its trials are, which questions have been well-explored, and which questions should be explored next.

Building on the PRECIS tool for trial design, I also showed how projecting a series of trials onto the same PRECIS graph could represent the degree to which a clinical research program has explored part of its problem space, specifically the space of pragmatic or explanatory design assumptions. These two problem spaces—the space of investigation types and questions across an entire research program and the space of trial design—can then aid judgments about whether shared results may be an artifact of certain pragmatic or explanatory design decisions. It can also reveal under-explored hypotheses or dimensions of trial design, and hence, inform decisions about the kinds of future studies and trial designs needed to achieve a robust result.

The aim of this chapter is to provide a more detailed illustration of how this model can be put to use in practice. In the ideal research program, a robust pattern of treatment efficacy and effectiveness is observed across all stages of research, from the pre-clinical—in vitro and in vivo—studies to the so-called phase I, II, and III studies in human subjects. A consistently positive pattern of results throughout this long process justifies both the epistemic and ethical requirements of continuing research into the particular treatment. Insofar as studies within and between the different stages adopt different experimental heuristics, then each successive, positive result should increase our confidence in the reality of the treatment's efficacy or effectiveness. This growing confidence in turn discharges the ethical demands of clinical equipoise, i.e., that before each trial, there exist a state of honest, professional uncertainty amongst the expert medical community as to the therapeutic merits of each arm in the study.

Perhaps unsurprisingly, such an ideal scenario is rare. Often there are discordant results: variable effect sizes, or even a variable direction of effect, across and between the stages of research; in other words, failures of consistency ("horizontal" robustness) or congruence ("vertical" robustness). Just as I have argued throughout this thesis, when these failures of robustness occur, thoughtful use of heuristics and meta-heuristics become more important than ever to guide the necessary changes.

The case I will examine here involves recent work on *moxifloxacin* (hereafter just "moxi") for the treatment of mycobacterium tuberculosis (hereafter just "TB"). Moxi is one of a family of antibiotics—the fluoroquinolones—that showed great promise in vitro and in murine models for reducing the treatment times of TB. Unfortunately, its effectiveness for similarly reducing treatment times in humans has been less promising. Five phase II studies with moxi, published between 2006 and 2009, produced discordant results: three studies showed modest, but statistically significant improvement over the standard TB treatment regimen; two studies showed no improvement.

Given that the robust pattern of effect from the earlier studies is now called into question, how should research with moxi proceed? Is the solution simply to do another study? Is a null result in two of the five studies enough to rule out further research? Are three positive results out of five enough to justify a phase III trial? Or is the discordance of the result a sign that there are more fundamental questions from the earlier stages of research that need to be re-addressed? The answers to these questions turn on heuristics and meta-heuristics, some of which are general scientific principles (e.g., "get more data"), others of which will be specific to an understanding of how different models of TB disease relate to one another (e.g., the kinds of results from mouse models of TB that translate well into human studies). The model of problem space is therefore only the beginning of the analysis I present here. It provides the sketch which the heuristic details must then fill in.

I begin in the next section by laying out the problem space for moxi research. Then, in §5.2, I discuss some alternative strategies for resolving the state of discordance. I argue that a statistical meta-analysis is unlikely to be informative, whereas a robustness analysis can reveal particular methodological hypotheses that warrant further investigation.

One such hypothesis involves the translation of evidence from animal models into human studies. This is an essential robustness consideration for many novel interventions, and yet it has received remarkably little attention to date. In §5.3, I discuss some of the recent attempts to establish more rigorous methodological measures for the quality of animal studies, as well as, identifiers of animal results that are likely to translate well into human trials.

I conclude in §5.4 with a discussion of what this case ultimately demonstrates about my model for a research program's problem space. Further work and development of the model is certainly needed; nevertheless, even this initial formulation can sharpen arguments about the evidential state of a research program; how best to perturb designs in order to make an optimal contribution to the field; and how to balance robustness considerations with the ethical demands of clinical equipoise.

5.1 The Moxifloxacin-TB Problem Space

TB is not a condition that receives much attention in the developed world, and yet, one-third of the world's population is infected with the disease. While only 5-10% of those infected will ever become sick or infectious themselves, TB is nevertheless a pressing global health concern. It accounted for 1.7 million deaths in 2009, as well as 23% of all deaths amongst people with HIV/AIDS. More troubling still is the increasing incidence of "multidrug-resistant" and "extensively drug-resistant" strains of TB, for which the standard treatment regimen is ineffective (World Health Organization 2010).

Fortunately, the majority of TB cases are treatable. The current standard treatment for drug-susceptible TB is a four-drug regimen (isoniazid, rifampin, pyrazinamide, and ethambutol) administered for six months. Since what ultimately matters is not just eliminating the active bacteria (which can happen quickly) but preventing future relapse, a large part of the challenge to eliminating TB is ensuring full compliance with this regimen for six months. Thus, the central aim of new research is to find a shorter or simpler regimen.

Unfortunately, evaluating the outcome of no-future-relapse is costly, requiring at least two years of patient follow-up. Such an undertaking is typically not possible until the phase III stage of research. As a result, most of the phase II studies with moxi have heuristically adopted eight-week culture conversion as a surrogate endpoint. This means that a patient whose sputum culture has converted to TB-negative within eight weeks of beginning treatment is considered a positive outcome.[1] Earlier stage studies adopt various other surrogate endpoints, depending

[1] Sputum is material expelled from the patient's lungs or collected from their saliva. If a patient has a bacterial infection, then this material will contain the microbacteria of interest. The sputum sample is collected (and possibly stored) by the researchers, so that the bacteria can be cultivated in a growth medium, either a solid agar medium (sometimes called a "plate") or a liquid broth medium. The amount of viable bacteria found in the sputum, usually measured in cfu/ml (colony forming units per milliliter), is a surrogate for the amount of bacteria

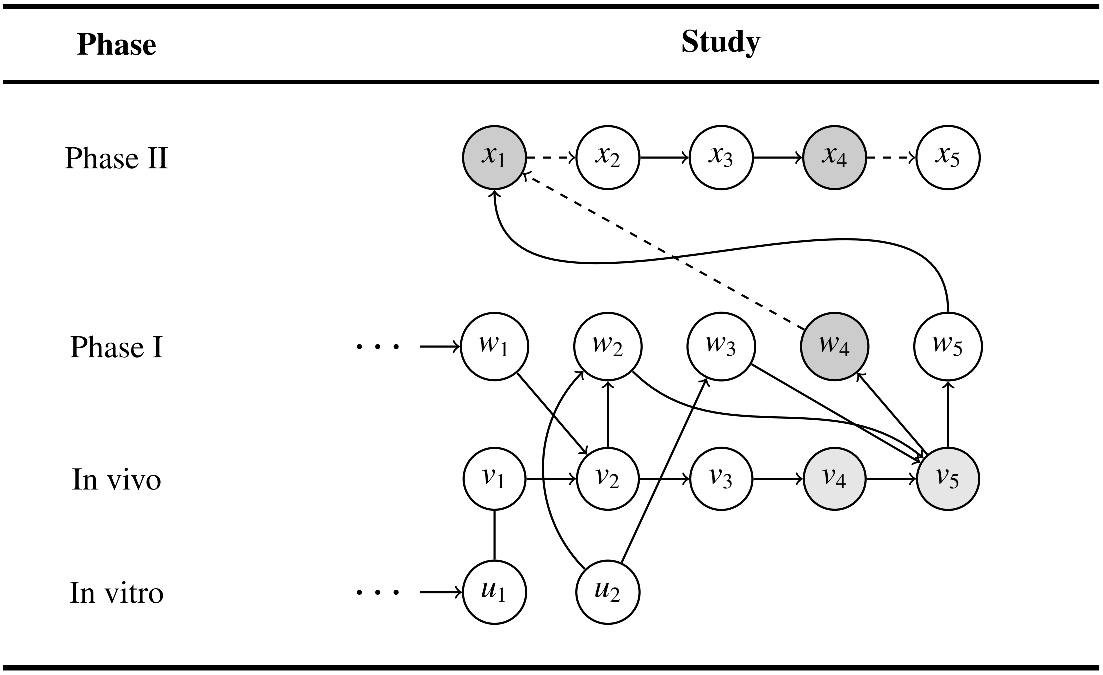

Figure 5.1: Model of moxifloxacin-TB problem space

upon the model in question. The specific endpoint used in each study will be described in the analyses below.

The fluoroquinolone family of drugs is a widely used and extensively tested class of antimicrobial agents. So much so, that moxi, a single member of this class, warranted its own supplement in the journal, *Clinical Infections Diseases*, in 2005—just five years after it was first approved for testing. In addition to analyses of the pharmacokinetics, pharmacodynamics, and safety, the articles in that volume discuss moxi's effectiveness for treating pneumonia, rhinosinusitis, and acute exacerbations of chronic obstructive pulmonary disease. It is against this background of wide use and effectiveness that the ensuing discussion of moxi's specific effectiveness for the treatment of TB must be set.

Figure 5.1 is a representation of the current state of moxi research, an instantiation of the problem space model introduced in §4.4. As before, the nodes in the figure each correspond to a study; the arrows between them representing epistemic relationships (as evidenced by the citations and discussion in each of the published study reports). Positive studies are represented by white nodes, null studies or studies with mixed results are represented by shaded nodes (a mixed result is a lighter shade than a null result). Epistemic relationships are also transitive, so a study connected further up the chain of research (e.g., x_1) is justified by the work of all the studies that came before it (e.g., w_5, v_5, v_4, w_3, u_2, etc.).

in the patient's lungs.

I will have much more to say about each of these 17 studies and their epistemic relationships (the impatient can immediately look to tables 5.1 and 5.2 for a summary), but before even getting into the details, it should be immediately clear how much messier is the picture of an actual research program as compared to the idealized picture presented in the last chapter. In particular, moxi's in vivo and phase I stages are "fluid", in the sense that in vitro studies provide the justification for phase I studies, which then justify the aims and structure of animal studies. This illustrates the somewhat arbitrary nature of the traditional phase divisions. It also speaks to the fact that the methodological progression of clinical trials in humans will not always be the same. Sometimes stages are repeated, sometimes they are skipped, depending upon the accumulating body of evidence and the demand for a novel intervention. This is a simple fact of clinical research, but it is nevertheless an important methodological point, made explicit in the representation of the model.

I should also comment on the narrowness of the problem space. As I noted above, there is much research into uses of moxi for conditions besides TB. There is also much TB research that has nothing to do with moxi. Both of these fields of research are still relevant to the model here, and important to consider when evaluating the quality of evidence in these studies. Nevertheless, trying to usefully represent all of the possibly relevant research in a single problem space would be too cumbersome, obscuring many of the immediate advantages. The picture I present here reflects some judgment about the studies that ought to be included in order to resolve the case of underdetermination at hand. But it is always important to bear in mind that this is only one small fraction of the total research. Just as I described, in the last chapter, how we might imagine "zooming in" on the individual nodes to reveal the details of a PRECIS graph for each study; so too can we imagine "zooming out" to reveal all the additional research dimensions: studies with all the other agents for TB, studies with moxi for the treatment of other conditions, etc.

5.1.1 Pre-Clinical and Phase I Moxifloxacin Research

Ji et al. (1998) was the first published, pre-clinical study with moxi that was specifically for the treatment of TB. It evaluated moxi alongside two other fluoroquinolones, clinafloxacin and sparfloxacin. It was a combination study, with an in vitro (represented as v_1 in fig. 5.1) and in vivo (u_1) experiment. The ellipses leading to u_1 in figure 5.1 represent the epistemic influence of the larger field of moxi and fluoroquinolone research.

As an in vitro study, Ji et al. was interested in the MIC, the "minimum inhibitory concentration" in which the TB bacteria's growth, using a solid culture, is reduced by 99%. They found the necessary MIC of all three drugs to be similar, although moxi and sparfloxacin performed

Table 5.1: Summary of pre-clinical moxifloxacin-TB studies

In vitro	**Key**	**Result**		**Media**	**Measure**	**Comparison**
Ji 1998	u_1	Positive		Solid	MIC	Sparfloxacin, clinafloxacin
Gillespie 1999	u_2	Positive		Solid	MIC	Isoniazid, ciprofloxacin, levofloxacin, ofloxacin, sparfloxacin
In vivo			**Mouse (N)**			
Ji 1998	v_1	Positive	Swiss (360)	Solid	30 day status	Sparfloxacin, clinafloxacin
Miyazaki 1999	v_2	Positive	Swiss-Webster (48)	Solid	7 wk. status	Moxi plus isoniazid
Lounis 2001	v_3	Positive	Swiss (380)	Solid	6 mo. status	Standard, streptomycin
Yoshimatsu 2002	v_4	Mixed	BALB/c (60)	Solid	3-day, 4 wk. status	Isoniazid, various moxi dosages
Nueremberger 2004	v_5	Mixed	BALB/c (270)	Solid	6 mo. status	Standard, pyrazinamide, isoniazid, rifampin

slightly better than did clinafloxacin.

Their in vivo study analyzed 360 Swiss mice (plus 30 that were sacrificed immediately to serve as a baseline), randomized amongst 12 different treatment groups: (1) isoniazid; (2-7) moxi and clinafloxacin, each given at three different dosages; (8-11) sparfloxacin given at four different doages; and (12) the untreated control. All of the isoniazid arm, and nearly all of the mice in the sparfloxacin arms survived to 30 days (where death at 15 days was the baseline for TB infection in these mice). Only seven mice in the moxi ams died, but since all of these deaths occurred between day 7 and 10 the authors’ argue that they were likely not due to the effects of TB. In contrast, all of the mice in the clinafloxacin arms did worse than the control. Ji et al. concluded that sparfloxacin and moxi were promising agents for future treatment of TB and further study was warranted.

Stass et al. (1998) was an early phase I, safety, tolerability, and pharmacokinetic (PK) study of moxi, conducted at a single site in Germany (w_1). As above, the ellipses leading to this study in figure 5.1 represents the larger field of fluorquinolone research. Although this was a study done with healthy volunteers, and therefore not a test of moxi for TB, I have included here because it was an often cited study for subsequent work.

They ran a six arm, randomized study with 45 healthy, male volunteers, each arm receiving

Table 5.2: Summary of phase I and II moxifloxacin-TB studies

Phase I	Key	Result	Sites (N)	Media	Measure	Comparison
Stass 1998	w_1	Positive	Germany (45)		2-day PK	Placebo, 6 moxi dosages
Gosling 2003	w_2	Positive	Tanzania (43)	Solid	5-day cfu/ml	Isoniazid, rifampin
Pletz 2004	w_3	Positive	Germany (17)		5-day cfu/ml	Isoniazid
Gillespie 2005	w_4	Null	Tanzania (13)	Solid	5-day cfu/ml	Historical monotherapy
Johnson 2005	w_5	Positive	Brazil (40)	Solid	2-day, 5-day cfu/ml	Isonizaid, levofloxacin, gatifloxacin
Phase II						
Burman 2006	x_1	Null	Uganda (175), United States (102)	Liquid, solid	8 wk. neg.	Ethambutol
Rustomjee 2007	x_2	Positive	South Africa (107)	Liquid, solid*	Rate to neg.	Ethambutol
Conde 2009	x_3	Positive	Brazil (125)	Solid only	8 wk. neg.	Ethambutol
Dorman 2009	x_4	Null	Uganda (182), United States (84), South Africa (31), Brazil (19), Spain (12)	Liquid, solid	8 wk. neg.	Isoniazid
Wang 2009**	x_5	Positive	Taiwan (123)	Liquid, solid	6 wk. neg.	Standard

*The significant, positive result was only found on the solid media.
**An unrandomized and unblinded study.

a different dosage of moxi (from 50mg up to 800mg). Blood, saliva, and urine samples were then collected over a 48-hour period (or a 96-hour period for the 800 mg group). They found moxi's pharmacokinetics to be "favorable", noting "good absorption and distribution, concentrations above the MICs for relevant microorganisms in plasma, saliva, and urine, and a long terminal half-life controlling the level of exposure to the drug" and an "excellent safety profile" (Stass et al. 1998, p.2064).

Gillespie et al. (1999) was another in vitro study (u_2). They evaluated moxi's activity, along with that of five other antimicrobial agents (isoniazid, ciprofloxacin, levofloxacin, ofloxacin, and sparfloxacin), against TB (and three other mycobacteria) on a solid medium. Their result was largely consistent with Ji et al., finding moxi to be the most promising of the tested agents for treating TB. They also noted that its in vitro properties were similar to isoniazid, another sterilizing drug already in the standard treatment regimen.

Miyazaki et al. (1999) was a second test of moxi in a mouse model of TB (v_2). They analyzed 48 Swiss-Webster mice, randomized into three treatment arms of 16 mice per arm: (1) a control arm given nothing but water; (2) a moxi arm; and (3) a moxi plus isoniazid arm. All of the mice in the moxi and moxi plus isoniazid arms survived to seven weeks, as opposed to about half of the mice in the control arm.[2] They argued that further in vivo work with moxi was warranted. They also noted moxi's similar performance to isoniazid in mice—a congruence between in vitro and in vivo studies—and argued that this was a sign of positive potential for its use in multi-drug treatment of human TB.

Lounis et al. (2001) was the first mouse model to evaluate moxi within variations of the standard multi-drug regimen (v_3). They analyzed 380 Swiss mice, randomized into eight treatment groups of 40 to 60 mice each (except for group 2, which only had 10 mice), ranging from an untreated control to the standard, six-month daily regimen, to various substitutions of moxi and streptomycin into the standard, to a once-weekly regimen of the streptomycin, isoniazid, rifapentine, and moxi for six months. Their findings were largely positive. All the mice treated in the moxi arms of the trial survived to six months (less those that were sacrificed for analysis at earlier time points); whereas two of the remaining mice in the streptomycin arms were still culture positive after six months. The subsequent three-month relapse rates also favored moxi, at 15% versus 61% for streptomycin. The addition of moxi to the six-month weekly dosage was also only marginally inferior to the six-month daily standard, which lead Lounis et al. to cautiously conclude that moxi might have potential for simplifying the standard regimen.

Yoshimatsu et al. (2002) was a dosage test of moxi as a monotherapy in the mouse model

[2] The difference in the expected lifespan of infected mice in Miyazaki et al.'s model versus Ji et al.'s model is striking. Although Miyazaki et al. do cite Ji et al.'s work as part of the background for their investigation, they offer no discussion or comment on this difference. Whether this is a valuable perturbation for the sake of robustness, a methodological oversight, or just an irrelevant detail is unclear.

(v_4). They analyzed 100 BALB/c mice, randomized into ten treatment groups (20 additional mice were sacrificed after infection to provide the baseline): (1) the untreated control; (2) daily isoniazid at 25 mg/kg of body weight; (3-6) daily moxi at 25 mg/kg, (7-10) weekly moxi at 50, 100, 200, and 400 mg/kg. After one treatment, five mice from the isoniazid and each of the weekly dosage groups (i.e., 1, 7-10) were sacrificed to measure the effect of a single treatment. After three days of treatment, half of the mice from each of the daily dosage groups (i.e., 3-6) were sacrificed. After four weeks of treatment, all of the remaining mice were sacrificed for the final analysis. None of the mice (including those in the control arm) died from TB during the course of the study. While low dosages of both isoniazid and moxi were not significantly better than the controls; the 200 and 400 mg/kg dosages of moxi did show significant improvement in lowering the CFU counts ("colony-forming unit", a measure of the active bacteria) by the three-day mark.

Discordant with Lounis et al.'s result of the weekly moxi-containing regimen's efficacy, Yoshimatsu et al. did not see any significant bactericidal effect from the weekly dosages of moxi. Given that Yoshimatsu et al.'s was a test of moxi as monotherapy, they suggest that Lounis et al.'s earlier result could have been due to interaction effects with moxi and the other drugs in the regimen. Yoshimatsu et al. also note some "untoward" side effects from higher dosages of moxi: failure to gain weight, decreased activity, and unkempt fur after four weeks. Thus, they concluded that high daily dosages of moxi were promising, but further studies of weekly combination therapies with moxi, as well as its toxicity effects on healthy mice, were still needed.

Gosling et al. (2003) was a phase I, early-bactericidal activity (EBA) test of moxi (w_2). They analyzed 43 pulmonary TB patients from a single hospital in Tanzania, randomized into three treatment groups: (1) daily isoniazid at 300 mg; (2) daily rifampin at 600 mg; and (3) daily moxi at 400 mg. They collected daily sputum samples for five days, and analyzed them for the reduction in CFU counts using solid growth media. Their results were concordant with the in vitro and in vivo results of moxi's bactericidal activity, again finding moxi to be similar in activity to isoniazid. By this time, they felt ready to conclude that "[c]linical trials to determine whether regimens containing moxifloxacin bring higher rates of culture conversion at two months should be performed as soon as sufficient safety data are available" (Gosling et al. 2003, p.1345).

Nueremberger et al. (2004) was another murine model, similar to Lounis et al., testing moxi's potential as part of the standard, multi-drug regimen for six months (v_5). They analyzed 270 BALB/c mice, randomized into six treatment groups: (1) a positive control receiving the standard multi-drug regimen of rifampin, isoniazid, and pyrazinamide given daily for two months and then daily rifampin and isoniazid for another four months; (2) adding daily moxi to the standard for both the two month and four month periods; (3) replacing pyrazinamide with moxi in the standard regimen; (4) replacing isoniazid with moxi in the standard; (5) replacing rifampin with moxi in the standard; and (6) an untreated, negative control. Their findings were "patently clear":

> The addition of [moxi] to the standard regimen resulted in a modest, but significant, improvement in the bactericidal activity at two and three months but did not shorten the time to culture conversion by even one month. These experimental results support the rationale for ongoing randomized clinical trials designed to test whether the addition of [moxi] to the standard regimen will increase the proportion of patients with negative sputum cultures after two months of therapy. However, it is unclear that the addition of [moxi] will permit shortening of the duration of therapy. Rather, it was the substitution of [moxi] for [isoniazid] in the standard regimen that resulted in a dramatic increase in potency (Nueremberger et al. 2004, p.424).[3]

Nueremberger et al.'s conclusion is thus mixed. Since reduction in overall treatment time is one of the major goals of TB research, their finding is disappointing. Nevertheless, their identification of a novel hypothesis— substituting moxi for isoniazid in the standard regimen will improve potency—can be considered a positive development. It is also the hypothesis the directly informs the design of Dorman et al.'s phase II study (to be discussed below).

Pletz et al. (2004) was a phase I, EBA study (w_3). They analyzed 17 patients from a single hospital in Germany, randomized to receive either (1) daily isoniazid at 6mg/kg of body weight or (2) daily moxi at 400 mg. They collected sputum samples before treatment, after 2 days, and after 5 days (although they do not describe their method and medium of analysis). Their result was largely consistent with Gosling et al.'s, finding moxi to be very similar to isoniazid.

Gillespie et al. (2005) was also a phase I, EBA study (w_4). They analyzed 13 patients from a single hospital in Tanzania, all given a combination of daily moxi (400 mg) and isoniazid (300 mg) for 5 days. At the end of treatment, they collected sputum samples and measured the CFU reduction in solid media. They found no significant improvement with the combination of moxi and isoniazid over monotherapeutic results from earlier studies at the same site. Yet, they acknowledged that their sample size was really too small to be conclusive.

Johnson et al. (2006) is the final phase I, EBA study (w_5) that I will mention here. They analyzed 40 patients from a single hospital in Brazil, randomized into 4 treatment groups: (1) daily isoniazid at 300 mg; (2) daily levofloxacin at 1000 mg; (3) daily gatifloxacin at 400 mg; and (4) daily moxi at 400 mg. Sputum was collected daily, beginning 2 days prior to treatment,

[3] It is not entirely clear to which study Nueremberger et al. are referring when they mention "ongoing randomized clinical trials", but it seems likely that they mean, at least, Burman et al. (2006), discussed below, which would have been underway by this time.

and continuing through seven days of treatment. The cultures were analyzed using solid media. They found only a minor improvement in EBA between isoniazid, moxi, and gatifloxacin.

Before moving on to discuss phase II trials, there are a number of points to emphasize about the pre-clinical and phase I studies. First, as I discussed at the end of the last chapter, there are a host of questions involved in the development and testing of any new medical intervention. This review of the early moxi-TB research shows not only how many questions get addressed from one study to the next, but also how many different questions are involved in each individual study. Three out of the five in vivo studies, for example, had six or more treatment arms, testing a range of potential drug dosages or substitutions. The question in those studies is not the binary, "is treatment *A* better than treatment *B*?", but the more open-ended, "which of these treatment options appears to be the most promising?" Given this, the summary label of "positive" study that I use in tables 5.1 and 5.2 should be read as a significant simplification.

Second, it is striking just how diverse most of these studies were. While there is overlap in certain design elements, e.g., the comparison between moxi and isoniazid persists throughout most of the studies, there was no mere repetition in any of the study questions. Subsequent research was also often built upon previous results, as is perhaps most evident in the standard adoption of the 400 mg dosage of moxi following Yoshimatsu et al.'s (2002) result. This variation of experimental design speaks very favorably to the potential robustness of the whole research program; lending credence to the belief that moxi's positive results are not artifactual.

Third, such a demonstration of robustness is an essential step for satisfying the ethical requirement of clinical equipoise. Despite all the perturbation of study questions and experimental designs to this point, one could certainly make the argument that there should exist disagreement about the preferred regimen for drug-susceptible TB. Despite some mixed conclusions coming in the later in vivo studies, there did seem to be an overall trend toward moxi's efficacy.

5.1.2 Phase II Moxifloxacin Research

Burman et al. (2006) was the first published phase II study (x_1). They analyzed 277 patients across multiple sites in the United States and Africa, randomized to either (1) the standard four-drug regimen or (2) the standard with moxi substituted for ethambutol. Sputum cultures were grown and analyzed using both liquid and solid media. Although moxi showed possible increased activity at earlier time points, it did not affect eight-week sputum culture status. They also stratified their analysis by continent, and found that African patients, despite the highest rate of compliance, responded far less to either treatment than did patients in the United States.

They concluded that further research with moxi was needed, but it seemed unlikely to shorten the overall treatment time for TB.

It is worth pausing here momentarily to reflect upon the choice of comparator drug in this study. Burman et al.'s decision to compare a moxi regimen to an ethambutol regimen is not unique. As I will soon show, two of the other four phase II studies also compared moxi's effectiveness as a substitution for ethambutol in the standard regimen. Yet, none of the earlier studies compared moxi to ethambutol. This reflects the use of a heuristic from ethambutol-TB research.

Ethambutol is the least active of the agents in the standard regimen (Dickinson et al. 1977, Steenwinkel et al. 2010) and its role in the regimen is to protect against rifampin resistance when there is already isoniazid resistance (Zhu et al. 2004). For patients with drug-susceptible TB (i.e., who do not have a resistance to rifampin and isoniazid), the heuristic is that ethambutol is not really needed (and sometimes it is excluded from the regimen altogether). Hence, the standard regimen could potentially be improved by replacing the least active agent, ethambutol, with the more active agent, moxi.

This speaks to the issue I raised above about how this problem space is connected to others. The heuristic behind the substitution of moxi for ethambutol is not justified on the basis of earlier work with moxi, but on the basis of earlier work with ethambutol. This is in contrast to Dorman et al.'s study (see below), whose hypothesis is based on earlier work with moxi in mouse models.

Rustomjee et al. (2007) was the next phase II study published (x_2). Theirs was a four-arm study comparing (1) the standard regimen versus three different flouroquinolone substitutions for ethambutol: (2) ofloxacin, (3) gatifloxacin, (4) and moxi. It was conducted at a single site in Durban, South Africa, and like Burman et al., evaluated sputum cultures with both liquid and solid media. However, unlike Burman et al., they did not use the eight-week culture conversion status as the surrogate endpoint, and instead, adopted the rate at which cultures converted to TB-negative.

They analyzed 217 patients in total (approximately 55 patients per arm) and found that both gatifloxacin and moxi improved the rate at which sputum cultures converted to TB-negative. As in Burman et al.'s study before them, neither moxi nor gatifloxacin showed any increased effect on eight-week sputum culture status. Nevertheless, because of their alternative endpoint, Rustomjee et al. take their result to support the opposite conclusion. They cite two articles on the treatment of TB-HIV/AIDS co-infection to argue that culture conversion rate is a better surrogate endpoint than the two month culture status. They also note that a significant difference between the moxi and control arms was only found with solid media. Whether despite or because of these breaks in methodology with Burman et al.'s previous work, they nevertheless conclude that a phase III trial with moxi is warranted.

Conde et al. (2009) was another single-site (Rio de Janeiro, Brazil), two-arm study comparing the standard regimen to the substitution of moxi for ethambutol (x_3). Like Burman et al., they used the eight-week conversion as the surrogate endpoint, but unlike the two earlier studies, they used only solid media cultures. Analyzing a total of 125 patients, they found a significant difference favoring the moxi arm.

Dorman et al. (2009) was another multi-site study, this time with locations in South Africa, Uganda, North America, Brazil, and Spain, which investigated the standard regimen versus a regimen where moxi was substituted for isoniazid (x_4). As I have noted, this shift from substituting for ethambutol was justified on the basis of results from mouse models (Nueremberger et al. 2004). Dorman et al. used two-month culture conversion status as their surrogate endpoint and evaluated cultures of both liquid and solid media.

After analyzing 328 patients, 213 of whom came from African sites (65%), they found no significant difference between the moxi and control arms. They were able to show that enrollment at an African site was associated with a lower likelihood of eight-week culture conversion status (regardless of the treatment received), but could not conclude that the prospects for reduced treatment time with moxi were positive.

The last TB study with moxi to be published (as of the time of my writing) was Wang et al. (2009) (x_5). Unlike all of the previous four studies, which were randomized and double-blind, this was an unblinded, non-randomized study; conducted at a single site in Taiwan. Rather than a substitution, they compared the standard four-drug regimen to the standard plus moxi. Cultures were analyzed on both liquid and solid media, but again, unlike any of the previous studies, the surrogate endpoint adopted was six-week (rather than the usual eight-week) culture conversion, on the grounds that "using [two-month culture conversion] alone as the primary endpoint does not reflect the entire spectrum of effectiveness of a fluoroquinolone-containing anti-tuberculosis regimen" (Wang et al. 2009, p.65). They analyzed 123 patients and found a significant improvement with the moxi arm at six-weeks.

Given this state of discordance—three positive phase II studies and two null studies—how should research on moxi for the treatment of drug-susceptible TB proceed? From a purely epistemic standpoint, this is a difficult case of underdetermination. Coming out of the pre-clinical and phase I stages of research, moxi looked reasonably robust and promising. There may have been some disappointing results, but nothing to obviously rule it out. Yet, the same cannot be said for the phase II studies. The failure to translate the success of moxi from the mouse models to the human studies is troubling. This is especially true in Dorman et al.'s study, given that the moxi for isoniazid substitution was the striking conclusion from Nueremberger et al.

Nevertheless, the discordance does not liberate us from the ability (thanks to good sense) or need (in light of the global health concern) to make a decision. Significant resources have already been spent on investigating moxi's potential for shortening the treatment time. Is the total evidence, as discordant or unclear as it may be, enough to justify spending further resources?

5.2 Alternative Analyses

One obvious approach to resolving the discordance would be to conduct more phase II studies. Assuming that each of these published studies is of high-quality and internally valid, we might conclude that although there is presently insufficient evidence for or against moxi, another high-quality trial has the potential to finally tip the balance one way or the other. Across scientific investigations generally, this is a reasonable and common heuristic, let us call it κ_1:

κ_1 : To resolve discordant results, go and collect more data.

However, in medical research, particularly at the stage of clinical trials, cost is a serious constraint on the possible number of investigations. Not only is there a cost to designing and conducting a trial, but there is also the opportunity cost—the time lost by investigators and research subjects in testing one course of treatment rather than another. Even a well-funded research program, capable of running many early phase studies, might be better off cutting its losses and abandoning an unpromising experimental treatment in order to pursue something else. In light of this, we can articulate a meta-heuristic, K_1, constraining κ_1:

K_1 : When the cost of additional research can be justified, employ κ_1.

The moxi dilemma turns, in part, on the applicability of K_1. With five phase II studies now complete, is the cost of another study justifiable? To think about what another phase II study might be able to show, it is worth reflecting on what the five extant studies have already shown. There are many different ways to go about this. For example, a coarse-grained description of the problem sets two null studies against three positive studies. So we might naïvely think that three studies should outweigh two, and therefore, we have good reason to think that moxi will be an improvement in effectiveness.

However, the fact that one of the positive moxi studies was non-randomized and unblind might suggest that it ought to be excluded from the tally. Not that the evidence provided by Wang et al. is irrelevant, but its methodological differences from the other four studies might suggest that, despite its being called a phase II study, it is really a different mode of investigation. Assuming then that the remaining four studies are all of sufficiently high-quality and methodological homogeneity, we are left with a 2-2 stalemate, so perhaps another phase II study like these four could finally tip the evidential scale.

But of course, this kind of coarse-grained labeling of "high-quality" studies and tallies of study conclusions is insufficient. Indeed, we might just as easily restrict our tally by surrogate endpoint, noting that amongst the eight-week culture conversion studies, the balance is 2-1 in favor of the null result. Or that head-to-head comparisons of a moxi regimen vs. an ethambutol regimen gives us 2-1 in favor of the positive result. But until one of these interpretative heuristics is justified by some deeper methodological argument, there is no prima facie reason to think that any one is better than the others.

Much the same can be said of a somewhat more fine-grained approach that looks at the sample sizes in each of the studies. We note that the two null studies are both significantly larger than any of the positive studies. The Uganda sites alone, from Burman et al. and Dorman et al., are both larger sample populations than any of the three positive studies. This might lead us to think that the positive studies, due to their smaller sample size, should carry less epistemic weight, and therefore, we do not have good reason to think moxi will be an improvement. But this approach will not do either. It is not necessarily the case that a larger study sample supports a stronger inference.

5.2.1 Statistical meta-analysis

A more sophisticated strategy for resolving apparent discordance is the use of statistical meta-analysis. On its face, the trials show no clear trend toward effectiveness or lack thereof, but perhaps by pooling together the data provided by each study, more decisive evidence could be detected.

A valid, statistical meta-analysis critically depends upon the homogeneity of the studies and their data. Studies with different populations, using different outcome measures, or different measurement techniques are much harder to pool together. This kind of heterogeneity between studies does not categorically rule out the usefulness of a meta-analysis, but it becomes dependent upon arguments for why the heterogeneity can be permitted. This concern is similar to Stegenga's (2009) worries, discussed back in chapter 2, about translations between different modes of evidence.

In the case of moxi, although there is some evidence of homogeneity across the studies, particularly in the measurement techniques (as discussed in the methods section of each publication), the discussion sections across the studies reveals an underlying heterogeneity in methodology. For example, Rustomjee et al. are critical of the binary outcome measure, "culture negative at some time *t*", and question whether or not this is the appropriate surrogate endpoint. Insofar as their critique is well-founded, it significantly complicates the structure of any meta-analysis, since this is the endpoint adopted by all four of the other studies. Employing a meta-analysis with the alternative, rate of conversion, endpoint would require that the data from the four other studies be re-analyzed. Since both Burman et al. and Dorman et al. sampled the culture status at intervals of two weeks (half as often as Rustomjee et al.), the requisite data may not even be available.

Rustomjee et al. are also critical of liquid media. Since the liquid media results of their study showed no significant difference between moxi and the control at eight weeks, they argue that solid media and the continuous endpoint of rate of conversion "may be a more useful method of assessment" (Rustomjee et al. 2007, p.135). But such a meta-analysis, excluding the liquid media results from all of the studies (excepting Conde et al. who did not use liquid media), begs the question. To test Rustomjee et al.'s suggestion with the extant data, we would have to assume that moxi is effective, and then go back to see if liquid media across the studies failed to indicate its effectiveness. This runs the risk of finding a significant result simply through data mining.

Finally, there are questions about the internal validity of the studies. In both Burman et al. and Dorman et al., there was a significant difference in treatment response between the patients in Africa and the patients elsewhere in the world. As statistical outliers, there is a temptation to exclude them from a meta-analysis. Perhaps African patients represent a relevant treatment sub-group that does not respond as well to moxi, and thus, they should be excluded from biasing moxi's effectiveness in the overall drug-susceptible TB population.

It is an important question as to why this difference was observed in African patients in the two null studies. But to exclude them in a meta-analysis again begs the question in favor of moxi's effectiveness. Dorman et al. are right to conclude cautiously that treatment at an African site in their study was "associated" with a worse outcome. It could very well be that a medically relevant sub-population for treating TB has been identified. More to the point, however, is the fact that TB is most prevalent in Africa: 30% of all new cases of TB are in Africa, as well as, 80% of TB-HIV co-infection. For the global effort toward treating and controlling TB, a treatment that does not work on African patients is ultimately of very little interest.[4]

This is all to argue that the heterogeneity across these studies precludes the usefulness of a statistical meta-analysis. Indeed, the opposing conclusions amongst the trialists appear closely associated with the heterogeneity in trial design, i.e., their use of different experimental heuristics. It is to these differences that we now turn our attention.

[4] MacKenzie et al. (2011) note that the observed difference between African and non-African outcomes in Dorman et al.'s study is not explained by baseline severity of the disease, HIV status, age, smoking, diabetes, or race.

5.2.2 Robustness analysis

While a statistical meta-analysis is weakened by heterogeneous trial features, a qualitative, robustness analysis is strengthened. The kinds of experiments or models where robustness fails can often be very informative about the system or experimental features on which a result critically depends.

But despite the discordance over the primary result (i.e., moxi's effectiveness for shortening TB treatment time), it is important to see how all five studies have still demonstrated at least one robust result: Treatment with moxi is associated with increased culture conversions at time points earlier than eight weeks. This is a consistent result—invariant across experiments at the same mode of investigation. The relevant mode here is phase II studies ($x_1 \dots x_5$), and all five of the phase II studies shared this positive result.

The moxi results also show some degree of *congruence*—invariance across experiments at different modes of investigation. Moxi's early bactericidal activity was observed in mouse models (e.g., v_5) and phase I studies (w_2, w_3, and w_5). The congruence of this property, moving from these early modes of investigation into the later modes of phase I and phase II studies is another robust result.

Where robustness finally fails (or is called into question) is the translation of improved outcomes in mouse model to shorter treatment times in human studies (i.e., $w_5 \rightarrow x_1$). Thus, we might begin our robustness analysis by questioning the external validity of the mouse model. And indeed, there may be good reason to doubt that a mouse model of TB is a reliable predictor for success in humans. Chest cavitation is a factor known to influence the effectiveness of TB treatment. Specifically the presence of chest cavities is inversely correlated with treatment effectiveness. In Burman et al., for example, 206 out of 277 patients (74%) had cavities, and these individuals were 40% less likely to convert to TB-negative after the end of therapy. Mice, however, do not get chest cavities.

The assumption that results from a mouse model of TB provide evidential justification for a human model of TB reflects a heuristic:

> κ_2 : A significant result in a mouse model of TB is likely to translate into a positive result in human TB studies.

Like all heuristics, there is no guarantee that the translation assumed in κ_2 will be smooth. Nor is there a guarantee that it will hold at all. Nevertheless, there are times when the translation has been smooth. As I have already noted, the early activity of moxi was robust across mouse and human studies. It is clear that moxi has this biological effect. It is unclear whether this effect represents an outcome of interest.

Two different methodological issues are thus intertwined here. There is a question about the appropriate surrogate endpoint, for which (roughly) two experimental heuristics are on offer: (1) Eight-week culture conversion; (2) rate of culture conversion. There is also a question about the evidential import of the pre-clinical mouse model for TB. If eight-week culture conversion is the better surrogate, then the incongruence of results from the mouse model to the human studies challenges the import of the mouse model. On the other hand, if the rate of culture conversion is the better surrogate, then the congruence of results from the mouse model to the human studies is a confirmation of the mouse model's evidential import.

As noted above, Rustomjee et al. justify their surrogate with an appeal to two articles on TB-HIV/AIDS co-infection. Burman et al. justify their selection of the eight-week culture conversion endpoint (as well as the power in their study) with reference to earlier research on pyrazinamide, one of the drugs in the standard TB regimen. Pyrazinamide's addition to the regimen shortened treatment times by three months and increased eight-week conversion rates by an average of 13% (Burman et al. 2006, p.332). It is not obvious that either of these justifications is sufficient to set a methodological standard for research with moxi. In fact, the radical underdetermination of how to proceed with moxi suggests that this basic methodological assumption needs to be better understood.

But this is not to accuse either study (or both) of poor design. Indeed, for designing early phase trials, pre-clinical studies and results from related interventions are the only sources of evidence. The use of alternative heuristics, and conflicts about which is preferable, merely speaks to a research program for which the meta-heuristics have yet to be articulated. Thus, a robustness analysis of these five trials does not resolve the question of what to do next. Instead, it illuminates the deeper methodological questions that need to resolved first.

5.3 Animal Models and Human Models

One such deeper question was alluded to above: Namely, the relationship between animal models and human models. Stepping back momentarily from the moxi case, the failure of evidence to translate from animal to human studies is a general issue with robustness in medical research that is only now becoming more widely recognized. As ubiquitous as the use of animal models is in pre-clinical research, it is remarkable that there has been so little research on the validity or strength of these translations.

Perel et al. (2007) is one of the first ever systematic reviews of concordance between animal and human studies. They analyze six interventions, and find variable results in the translation of effectiveness:

1. Corticosteroids for head injury showed a benefit in animal models, but no benefit and an

increase in mortality rate for humans.

2. Antifibrinolytics for haemorrhage was inconclusive in animal models, but showed a benefit in humans.

3. Thrombolysis for stroke was shown to be effective for both animals and humans.

4. Tirilazad for stroke was effective in animals, but associated with worse outcomes in humans.

5. Antenatal corticosteroids to prevent neonatal respiratory distress syndrome was effective in reducing stress and mortality in human studies, but its effects on mortality were inconclusive in animal studies.

6. Bisphosphonates to treat osteoporosis was effective in both animals and humans.

The fact that the second intervention—antifibrinolytics for treating haemorrhage—was inconclusive across animal studies, and yet effective in humans is a surprising finding. A happy accident perhaps, but it is troubling to think that a treatment could move on to human studies without first showing a robust pattern in pre-clinical work.

The first and fourth case, which show effectiveness in animals that is not borne out in humans, bear the most resemblance to the situation with moxi. Perel et al. note that in the case of corticosteroids "[t]he [17 animal studies] were, however, from one laboratory, had little evidence on adverse effects, and did not examine comorbidities" (p.199). This is in contrast to the 113 thrombolysis animal studies, which investigated different animals, age ranges, intervals after stroke, and took place in different laboratories (p.198).

Yet, despite the consistency across animal studies of thrombolysis, Perel et al. do still voice some concern:

> Thrombolysis with tissue plasminogen activator was effective in animal models of stroke and the results agreed with the clinical trials. The animal studies were of poor quality, however, with evidence of publication bias. Our evidence for concordance may therefore be biased. We found over 100 experiments, totaling more than 3000 animals. The pooled result was therefore precise although not necessarily valid (Perel et al. 2007, p.198).

This is, in part, an acknowledgment of the limitations of meta-analysis that I noted earlier, albeit at an earlier stage in the research program. But assuming that at least some portion of the 113 animal models were of high quality, Perel et al.'s observation that they perturbed elements of their study designs does provide evidence of robust consistency. The fact that the intervention was then shown to be effective in humans is evidence of robust congruence. This is an instance of the ideal for medical research: A positive result in pre-clinical studies is subsequently demonstrated in human studies.

Perel et al.'s final conclusions are mixed, and Van der Worp et al.'s (2010) more recent discussion of this issue echoes some some of their concerns:

> Although there is no direct evidence of a causal relationship, it is likely that the recurrent failure of apparently promising interventions to improve outcomes in clinical trials has in part been caused by inadequate internal and external validity of preclinical studies and publication bias favouring positive studies. On the basis of ample empirical evidence from clinical trials and some evidence from preclinical studies, we suggest that the testing of treatment strategies in animal models of disease and its reporting should adopt standards similar to those in the clinic to ensure that decision making is based on high-quality and unbiased data (Van der Worp et al. 2010, p.6)

Some of the "standards similar to those in the clinic" that they suggest are ones discussed in the previous chapter, e.g., a methodological quality checklist like the CONSORT statement, and they endorse trial registration as a check on publication bias.

However, Van der Worp et al. also make the following provocative claim:

> Not only should the disease or injury itself [as investigated in the animal model] reflect the condition in humans as much as possible, but age, sex, and comorbidities should also be modeled where possible. The investigators should justify their selection of the model and outcome measures. In turn, *human clinical trials should be designed to replicate, as far as is possible, the circumstances under which efficacy has been observed in animals* (Ibid., pp.6-7, emphasis added).

Much of this passage is entirely consistent with Perel et al. and reflects certain ideals of robustness as an aim of research. We want to improve the strength of translations from animal to human studies, so we should design and perturb our animal studies to achieve maximum robustness. The final sentence of the excerpt, however, reflects a much more controversial claim. Indeed, a strong reading of the prescription implies significant naïveté about what it is possible to control in clinical research. The environment, history, and life-cycle of research animals is so unlike that of human subjects, so tightly controlled and sterilized, that it is impossible to expect to replicate anything similar in human studies. And as the moxi case exemplifies, there may simply be biological differences between the animals and humans, e.g., chest cavitation, that cannot be translated.

A weaker reading of the prescription is more plausible, but raises the issues discussed in the last chapter about what kinds of clinical trials are needed. Designing clinical trials to resemble the conditions of pre-clinical studies may improve the strength of translations going from animals to humans, but it weakens the strength of translations going from clinical trials to clinical practice. This is the now familiar distinction between explanatory and pragmatic trials designs. Both kinds of trial designs (or design dimensions, cf. chapter 2) have their place in a sequence of relevant scientific questions.

Happily, we can have it both ways. We need not insist upon making as strong a claim as do Van der Worp et al. Instead, we can elaborate the picture of a robustness-directed clinical research program, calling for consistency and congruence across the pre-clinical studies, moving from in vitro to in vivo models. Once we have this robust evidence, we then want explanatory trials in early phase human studies, seeking to establish congruence of results from the pre-clinical studies. If congruence is extended into phases I and II, we can then justify a state of clinical equipoise and call for the more pragmatic, phase III trials to evaluate effectiveness in conditions similar to those of clinical practice.

Chapter Summary

The moxi case illustrates the utility of the model I set out in chapter 4 in a number of different ways. In the first place, the process of constructing a model, like figure 5.1, for any given field of research reflects what is (at least implicitly) going on during the design and development of every new study. A researcher (or research team) surveys their field, making judgments about the quality and relevance of other studies with an eye to identifying a worthwhile hypothesis for investigation. Similarly, as I noted in §5.1, when constructing a research field's problem space, judgment must be made about what studies to include. A balance must be struck between an overly narrow and overly broad representation. If it is too narrow, then it is likely to overlook important studies. If it is too broad, then it is likely to become cumbersome and difficult to see how all of the studies relate to each other (and what the next step ought to be).

For example, the hypothesis tested in three of the five phase II studies discussed here, comparing a moxi-containing regimen versus an ethambutol regimen, is not derived from the earlier phase work with moxi, but from other work on ethambutol and the activity of all the standard drugs in the TB regimen. Given this fact, one could argue that the model in figure 5.1 was really too narrrow and would be improved by adding an epistemic link between x_1 and other studies, "external" to the moxi-TB field. But the claim here has never been that figure 5.1 is the perfect representation of moxi research (although I would argue that it is at least a useful representation). Rather, the claim is that *applying* the tool, i.e., constructing the model, actually brings certain methodological judgments to the fore—judgments about the current state of research and an optimal next step.

Much of the discussion in this chapter has been focused on exactly that question: What should the next step be for moxi research? As I argued in chapter 4, a well-designed study will be one that attempts to draw as much as possible on other work in the field, perturbing their heuristics and assumptions so as to make an optimal epistemic contribution. The model of problem space also helps to accomplish this task by providing an immediate visual representation of the field, making it clear how many and what kinds of studies are already out there, and how these are related to one another. For example, many null studies at one phase of research should raise questions about the translation of results from earlier phases. How connected are all of the null studies to earlier phases? How robust were the positive results from earlier phases? Just as I argued was the case for the PRECIS graph or my proposed modification of it for biological modeling (cf. chapters 2 and 3), the immediate visual representation of the model helps to make these questions more concrete.

The failure to find a robust pattern of effect at the phase II stage of research raises questions about whether or not the right kind or a sufficient degree of robustness was present in earlier phases. Going back to investigate these questions may reveal further questions about our basic understanding of the relevant biology and pharmacology. One of the lessons here is that plowing ahead to phase III with moxi, or even just ruling it out for further investigation, misses the essential, underlying methodological issues. And we will not be able to adequately judge whether a state of clinical equipoise ought to exist until we have a better understanding of the methodological heuristics.

What is the purpose of a phase II trial and what are the correct criteria for moving to phase III? The answer to this question (or complex of questions) relies upon methodological heuristics and meta-heuristics specifying the aims of the research program and the appropriate measures and thresholds of evidence. One understanding of phase II trials is that they are largely tests of safety and efficacy, designed to give us preliminary evidence of effectiveness. Five studies show that moxi is efficacious and, at least, no worse than the standard regimen. If the role of the phase II is to rule out unsafe or unpromising treatments for further testing, then moxi has arguably passed this test, and warranted a phase III investigation.

But perhaps the threshold should be set higher. If the role of a phase II trial is to rule in promising treatments for further testing, the case for moxi is more difficult. The two largest studies, most resembling the structure of a future phase III trial, showed no significant improvement over the standard. Moreover, the poor performance of moxi (and the standard regimen) at the African sites is perhaps a sign that the science of TB treatment needs to be re-evaluated at a more basic level. As such, pursuing moxi in the context of a large, pragmatic study without first improving our understanding of the underlying biology and methodology might simply be a waste of resources.

As discussed in §5.2., Burman et al. and Rustomjee et al. are not disputing the role of a phase II trial per se, but rather the appropriate surrogate measure for studying TB. This speaks to the layered nature of methodological heuristics. At a basic level, there is a question about the role of phase II studies in general: Are they supposed to rule out or rule in new treatments? And how might this change depending upon the condition and intervention in question? As I noted earlier, in the face of a global health concern, underdetermination cannot trump the need for researchers to make a decision. If a meta-analysis would be dubious and robustness analysis points to more basic questions, then the straightforward decision for moxi might be to say, no, a state of clinical equipoise is not justified, and therefore, phase III trials should not proceed. But this may discount an additional, ethical dimension. If the need for new, shorter TB treatment is dire enough, it may be necessary to lower the threshold for phase II. In which case, the prudent decision might be to acknowledge the methodological issues, set them aside for the time being, and proceed with further tests of moxi's effectiveness.

This kind of epistemic-ethical negotiation speaks to another point I raised earlier about the cost of additional experiments. For sciences with relatively low cost and low stakes (whatever those might be), it is entirely reasonable to set a high bar for robustness and demand that any result is evaluated across a range of relevant perturbations. In the course of medical research, for example, we might propose a kind of sliding scale, where the requirements of robustness in the in vitro stage are quite high, and then gradually decrease in reflection of cost as we move to animal models, human studies, and finally to phase III trials. There may even be times, as was the case with the recent H1N1 vaccine, where rigorous clinical research is radically abbreviated in order to deal with a global health threat (cf. Fedson 2005).

In such extreme cases, robustness considerations at each stage of research are clearly less pressing, but they are no less relevant. The ethical demands of providing patients with care, even if that care has yet to be rigorously evaluated, can and should come first. But this does not mean that we should ignore underlying scientific questions.

The appropriate balance between the epistemic demands of robustness and ethical demands of patient care is another consideration that could be built into a more comprehensive model. For example, we might take the necessity of finding new, effective treatments for African populations as grounds to set a minimum threshold of three positive, phase II studies at African sites before moving on to phase III. Such a prescription, driven by both empirical and ethical considerations, would dramatically change the structure of the problem space. Trials like Conde et al. and Wang et al. would still be informative about the efficacy of new treatments, but they would no longer figure into the justification for moving on to phase III. As a result, these non-African phase II trials could be given their own “line” in the model, between phase I and the African phase II trials. The outcome difference between African and non-African patients observed by Burman et al. and Dorman et al. would also become a more serious concern. Since all new treatments *must* show signs of success in Africa, this would place a greater emphasis on understanding the causes of this difference.

By representing the state of research, building in quality and relevance considerations, and leaving room for ethical and epistemic negotiation, the general modeling strategy I am advocating is a tool for scientists as much as it is for philosophers. For scientists, it can help to sharpen thinking about past and future research, addressing questions about what has been shown and what needs to be shown next. For philosophers, it sharpens our reflections about robustness considerations and how to optimally organize grand research projects, helping us to think about how scientists can best amalgamate evidence as they move through the stages of a research program.

Chapter 6

A Philosophy of Scientific Judgment

The investigations in this thesis have been wide-ranging—beginning with Pierre Duhem's ideas about physical experiments in the early 20th century, moving through the group selection controversy in biology in the 1970s and 1980s, and ending with an analysis of current clinical trials for tuberculosis treatments. In this final chapter, I retrace the main arguments and conclusions, emphasizing the intellectual threads that tie all of these topics together. I will then conclude with a discussion of how I see the project developing from this point forward.

6.1 Solving the Problem of Undetermination

Taking Duhem's philosophy as my starting point is misleading in some ways. At its core, this has not been a historical work. While I do think my hierarchical account of heuristics and meta-heuristics comports well with his descriptions of good sense, I have not been concerned to restrict my account of good sense to what he might have found acceptable. I make no claim that he had something like my view in mind or that he would have endorsed my view. Thankfully, my arguments for the methodological importance of heuristics, meta-heuristics, and the construction of problem space stand independently.

That said, Duhem's presence in the introductory chapter is not arbitrary. It is justified in two respects: First, he articulates the project of this thesis: How can we make good sense "more lucid and more vigilant"? If logic is not the rule, then how do scientists make decisions? How can we normatively evaluate these decisions? As familiar as the problem of underdetermination is for philosophers of science, there have been remarkably few (and remarkably feeble) attempts to develop a rigorous account of the solution. The fact that Duhem himself gestures towards a possible solution makes for a provocative starting point, since the seeds of a solution were sown alongside the very problem.

Second, the typical, unproductive responses to Duhem's problem, e.g., that scientific judgments are ungoverned and irrational or to be explained sociologically rather than philosophically, would have us miss out on all of the rich methodological subject-matter that I have discussed. Therefore, beginning with Duhem's ideas allows me to set up a poignant contrast between the uninteresting question about whether Duhem's problem can be solved (since it is obviously solved all of the time in scientific practice) and the far more interesting question about *how* scientist deal with underdetermination. What are the rules that guide their judgments? What implications does this have for understanding what is "good sense"? And what might philosophers say about this?

My approach to these problems has been to leverage work on scientific heuristics. Scientists are not helpless in the face of underdetermination. This is because they rely on rough-and-ready rules that tell them how to conceptualize the system of interest, what to assume in their models, and what to revise when things do not work as expected. Scientists do not think of their theories as sets of sentences, whose probabilities are adjusted in the face of evidence. They do not perform huge numbers of calculations to determine the proper course of action. They learn which heuristics to adopt, simplifying the problem in front of them, and then "acting" upon this simplified system. It is this fact about scientific practice that inspires the *philosophical* meta-heuristic driving my whole analysis: *The solution to any case of scientific underdetermination can be found in the (possibly implicit) use of heuristics.*

Like all meta-heuristics, mine is just one philosophical strategy from amongst many. It is not guaranteed to always work, and it will bias the analysis in systematic ways. But as the strategy of this entire thesis, its potential fruitfulness should now be beyond doubt. Focusing on the actual solutions to cases of underdetermination (rather than rational reconstructions)—working out the different heuristics employed by different researchers—has provided insight into methodological conflicts across the sciences. In physics, I showed a contrast in variable reduction heuristics versus model-world matching heuristics; each appropriate for different kinds of explanations. In emergency medicine, I showed how the use of fast-and-frugal heuristics was superior to logistic regression models for improving physician behavior. In biology, the over-use of context simplification and functional localization heuristics gave rise to the need for improved explanatory meta-heuristics, like the pathways of explanation, to make the relationship between kinds of experimental environments and inferences explicit. In ecology, the failure of the optimization approach, with its heuristics for static modeling of ecosystems, lead to the resilience approach, with its dynamic heuristics for the adaptive cycle. In clinical trials, I showed a distinction in applying the control heuristic, where limiting controls enabled higher patient accrual rates and favored pragmatic studies and effectiveness questions.

These cases demonstrate the ubiquity and diversity of heuristics, as well as, the applicability of my meta-heuristic account of good sense. Behind the application of any first-order heuristic judgment, there is a meta-heuristic that prescribes the conditions of the first-order heuristic's appropriate use. By organizing these heuristics and meta-heuristics into a picture of the relevant problem space, we make the methodological prescriptions explicit, and can then subject them to a more careful analysis. For example: "Does the meta-heuristic strategy coincide with the heuristics being applied?" The PRECIS graph of clinical trial design addresses exactly this kind of question. It provides a picture of the problem space that allows us to, first of all, judge whether a trial's design orientation (i.e., its character on the pragmatic-explanatory continuum) matches its stated research aims; and second, to articulate a pair of governing meta-heuristics: (1) A research strategy that calls for effectiveness questions needs to be matched by a trial that adopts pragmatic design heuristics; (2) a research strategy for efficacy or biological questions needs to be matched by a trial that adopts explanatory heuristics.

These two meta-heuristic prescriptions are implicit in the Thorpe et al.'s justification for the PRECIS tool, but by making them explicit, we draw out the connection to the first-order scientific heuristics of trial design. Moreover, we can then consider how these prescriptions and their representation of problem space might be used in other sciences. Indeed, it was not a far stretch to see how a similar graph could be useful for understanding models in biology. Utilizing Levins' dimensions of modeling, we could translate Griesemer and Wade's prescriptive pathways of explanation into meta-heuristics about the space of biological modeling and experimentation: Insofar as models neglect the dimension of realism, their inferences to natural explanations are weakened. While this is perhaps an obvious point in retrospect, it was nevertheless an interpretive heuristic overlooked by many group selection theorists.

Which brings me to the prospective advantages of focusing on the solutions to underdetermination. Elucidating heuristics, meta-heuristics, and their appropriate domains in problem space provides a more richly informative foundation on which new research can proceed. A biologist who understands how past models and experiments have explored the relevant problem space; who can compare the alternative meta-heuristic strategies given the system of interest, is far better equipped to design a novel model or experiment that will contribute to the overall understanding in their field. If the majority of prior models have made unrealistic assumptions, then there will be an obvious hole in the problem space, and a clear need to investigate this empty region.

Similarly, the clinical trialist who understands the PRECIS tool; who has a picture of how past trials have investigated the experimental intervention, is better equipped to ask a clinically relevant question and design the right trial for answering it. If the majority of prior studies have been highly explanatory, this fact will be obvious and underscore the need to design and run a pragmatic study, or vice-versa.

The analogy between pragmatic clinical trials and realistic biological models provides a relatively smooth translation of meta-heuristics from one domain to another. But there is little reason to suspect that, as a general rule, heuristics or meta-heuristics tailored to one domain will be well-suited for a different domain. The important heuristics for understanding each of the case-studies emerged from the insights of scientists or philosophers working in the specific domain, and it is the justification of a heuristic *for its own domain* that guides reflection at the meta-heuristic level. My account of good sense is thus a bottom-up analysis. What works for amalgamating evidence across a series of clinical trials need not work for amalgamating evidence in physics. In this sense, my account is highly pragmatic. The domain of applicability for a heuristic can be quite limited. What matters is that we understand how it works in its domain.

Like the heuristics that are its subject-matter, the lessons of my account are, for the most part, conditional and subject to revision. The use of heuristics, the need for meta-heuristics, and the organization of these within a problem space are the three universal features of my account. The appropriate content of these features will always be domain specific. I do not need to claim that the heuristics I have identified and labeled throughout this thesis are the best possible formulations of the judgment strategies and techniques. Nor do I need to claim that my representations of the various problem spaces are maximally useful or illuminating. These could all be quite wrong. The only essential point is the value of the philosophical strategy itself; how the attempt to explicate good sense with heuristics is revealing of scientific methodology and how to improve it.

6.2 The Strategy of Robustness

As powerful as heuristics are, they are also error-prone. The check against errors produced by a single heuristic or strategy is a search for robustness, ensuring that the important results are genuine and not artifacts of simplification. This is what I have called the *strategy of robustness*: A systematic perturbation of heuristics across the space of possible investigations. Results that are invariant under this perturbation have the *property of robustness*, giving us good reason to accept them as reliable.

While scientists do not need to be told to adopt this strategy—the use of multi-modal tests for instrument calibration or evidence evaluation is standard practice in just about any scientific field—philosophical analysis of robustness techniques and arguments has tended to make a couple of mistakes. First, the property of robustness has often been conflated with stability properties. The latter are properties of mathematical results, claims internal to an equation or system of equations. While I have no serious quarrel with mathematicians and scientists using the word "robust" to talk about such results, the kind of multi-modal robustness that is the result of using multiple techniques, experiments, or diverse heuristics is something quite different.

There are no quantitative methods that can demonstrate dependency relationships across, say, an in vitro study and a human clinical trial. And yet, intuitively, a result concordant across such diverse investigations still tells us something important. I have argued that this "intuitive" concordance is really the proper domain of robustness properties. Robust results are invariant under perturbation to heuristics. The experimental heuristics of an in vitro study are going to be quite different in kind from those in a human clinical trial. Insofar as a result (which in this case would be a direction of effect) is congruent across these changes, that is evidence that the result is not an artifact of the biases unique to each investigation.

The second common philosopher's mistake has been to focus exclusively on the property of robustness, overlooking its role as a strategy. Given that there are no universal, quantitative methods to combine the sometimes radically different kinds of data used to demonstrate a robust result, philosophers have expressed worries about the utility of the concept altogether. How can we make robustness arguments if we lack a universal way to distinguish robustness from pseudo-robustness?

Such a worry is not to be dismissed out of hand. Indeed, it is important to scrutinize results that seem to be robust, analyzing the heuristics underlying the different modes of investigation. Upon closer inspection, it may turn out that a result is less robust (or not robust at all), due to a lack of diversity in the sampled heuristics. The right way to go about exploring a problem space, and the right way to gather and amalgamate diverse data from different kinds of studies is going to be domain specific. But whatever the challenges posed by the analysis of some particular robustness argument, the general meta-heuristic strategy that instructs scientists to perturb their heuristics toward a sufficiently diverse set remains valid.

My analysis of the group selection controversy was largely devoted to unpacking the domain-specific heuristics and meta-heuristics needed to distinguish robustness from pseudo-robustness in that case. The orthodox view that group selection was largely inefficacious was, it turned out, dependent upon an array of inappropriate heuristics: a research heuristic prioritizing mathematical results over empirical investigations; a conceptual heuristic of descriptive localization; and a modeling heuristic of context simplification. All of these were so widely adopted amongst the community of researchers that a shared bias went undetected, leading to an under-described and under-explored problem space. The orthodox view thus *appeared* to have the requisite diversity to support a robustness claim.

Eventually, further theoretical and empirical work allowed for a better articulation of group selection's problem space. Wimsatt described the necessity of representing multiple levels of organization. Griesemer and Wade emphasized the importance of having more realistic biological models, as well as, different modes of investigations (i.e., laboratory models and field studies). Once these observations are used to re-describe the problem space, it becomes immediately clear that the orthodox view was unjustified and assumed a research domain that was far too narrow. Natural selection is ultimately about organisms in the wild, not variable dynamics in a mathematical system or hypothetical agents in a computer simulation. A solution in these computational domains that is not shown to be robust in more natural systems is not really a solution at all.

The story of assay sensitivity in the clinical research literature was similar. The assumption that a single trial, so long as it possessed assay sensitivity, was sufficient to establish the efficacy or effectiveness of a novel, medical intervention required a similarly skewed picture of the relevant research domain. Just as no single mathematical model (or even a related family of models) is sufficient to establish an explanation of real organisms in a natural system, so it is that no single clinical trial is enough to robustly establish that an intervention is effective. The idea that assay sensitivity replaces the need to interpret of a trial's result in the light of other study results is misguided at best. A single trial is always one investigation from amongst many that make up the complex epistemic network of a clinical research program.

Connecting back to Duhem's ideas on underdetermination, no experiment is a fully self-contained enterprise. Every experiment demands the use of good sense for its construction, execution, and interpretation. And each of these steps can be best informed if we understand where the particular experiment fits into its larger epistemic context. The more we understand about the whole research program, what has already been shown and what has yet to be shown, the better situated we are for making the next step. For clinical trials, this means a complete reversal of Temple and Ellenberg's view. A trial should be designed to be as un-self-contained as possible, drawing on the evidence and work of as many other high-quality and relevant studies as possible. In doing so, this will ensure that its results do more than just duplicate earlier work. A well-designed and well-executed study will ensure an optimal contribution to its field, providing either a demonstration of consistent and congruent robustness or suggesting limitations on the robustness of earlier evidence.

6.3 Modeling a Research Program's Problem Space

In the last two chapters, I moved from thinking about a problem space in the context of a single hypothesis to the problem space of multiple hypotheses across an entire research program. Robustness is at the center of the epistemic challenges facing a grand research program, such as the moxi-TB trials. The failures of robustness emerging at the phase II stage make this point more obvious, but even before these results were known, TB trialists have to interpret and ultimately build upon evidence from a vast array of different kinds of experiments. Which endpoint is the best surrogate for stable cure? Which kinds of mouse models provide the most reliable translations? What explains the observed difference in African versus non-African patient outcomes? None of these questions have definitive answers. The ongoing methodological discussion in that field reveals that they are still sorting out the right heuristics for the job, still working out what constitutes good sense.

The model of a clinical research program's problem space that I proposed in chapter 4, and then applied to the moxi studies in chapter 5, attempts to contribute to this methodological discussion in much the same way as the PRECIS graph contributes to the methodology of trial design. The idea is to make the "character" of a research program explicit: What are the different hypotheses being tested? Which results are robust? What is the next logical step? The network model is not a substitute for detailed answers to these questions, but it is a visual aid. One can simply look at the state of the model and tell how many studies have been done at each stage of research; whether or not these studies have drawn on the results of previous work; and whether or not the overall trend has been toward a positive result.

But beyond those factual features of a research program, there are also a host of judgments one can make. Are there enough studies being done at each stage? Or are there too many? Are there studies whose results should be more influential? Are there studies whose results are too influential? Is there robust evidence supporting the overall trend or has it just been a series of similar studies replicating essentially the same result? Making these judgments about the quality of research requires an appeal to heuristics or meta-heuristics, critiquing the design of individual studies, interpreting the importance of their results, and prescribing the appropriate thresholds for robustness within and between each stage.

6.4 Future Work

More thorough analysis of various scientific problem spaces is perhaps the most obvious extension of the ideas in this dissertation. The examples from physics, emergency medicine, and ecology that I touched on in the first chapter could all be given a more complete treatment, contrasting more of the relevant heuristics and meta-heuristic strategies. Even the case-studies of the later chapters, i.e., group selection models and moxifloxacin-TB research, are still only partially described. A single chapter, while adequate to illustrate the philosophical approach, is really insufficient to describe and organize all of the techniques and judgments that govern even a single research program. Yet, as this is the depth required to realize the full promise of the meta-heuristic framework, it remains an important task for elucidating the content of good sense in different domains.

The project of thoroughly characterizing a problem space also goes hand-in-hand with the project of improving the problem space representations themselves. I have relied on radar graphs, tree diagrams, and 2-dimensional network models, and while these are all inspired by extant methodological tools, they are also limited by the mode of presentation in this thesis. Black and white images on a piece of paper is certainly not the only visual medium available for organizing and analyzing research data. Indeed, it is easy to imagine how these representational tools could be developed into interactive software, permitting users to encode and adjust information with the click of a mouse. For example, "zooming in" on nodes in the network graph to reveal a trial's PRECIS graph need not be merely an appeal to the imagination. Future work in this direction therefore suggests opportunities to collaborate not only with scientists, but also with software engineers toward the development of more sophisticated and useful problem-space-modeling tools.

But in addition to these largely epistemic projects, there are also clear implications for policy work. In chapter 4, I largely set aside the project of making FDA policy recommendations, since the main focus of that chapter was to elucidate how the tools of good sense could be deployed in the context of clinical research (replacing the confusing guidelines surrounding assay sensitivity). But the conclusions of that chapter highlighted a number of ways in which the current regulatory guidelines are inadequate. Insofar as the FDA and the ICH E10 help to set the standards for clinical trials, it is important that they represent the epistemology of science accurately, and as such, the network model of a clinical research program's problem space provides a way to make some of their policies more concrete. For example, if we can show a robust pattern of effect throughout the history of the current standard treatment, then regulators can judge a PCT to be uninformative and unnecessary. On the other hand, if we model the history of the current standard treatment and come to see that it has not been robustly effective, then that would be good grounds for arguing that a placebo ought to be used for the evaluation of a novel intervention.

Beyond regulatory policy, there is more work to be done in developing general research policies for how to optimally structure a clinical research program. Chapter 5 really only begins to scratch the surface of this issue. The moxi case is a perfect example for this thesis, since it is an instance of underdetermination by discordant evidence and raises important questions about robustness. But just as important as understanding how to handle failures of robustness is understanding cases where robustness has been successfully achieved. Future work in this direction would contrast cases like moxi-TB with the successes in clinical research; searching for the heuristics that have worked in the past, figuring out why they worked, and when they are likely to fail. Such a project is not one for philosophers of science alone, but calls for intimate collaboration with the trialists and biologists working in health research.

Which brings me to the concluding note: The work in this dissertation lays a foundation for a new philosophical approach; a new way to analyze scientific research. It is a move away from rational reconstruction and a move toward better appreciation and understanding of the techniques and strategies that scientists actually use. Therefore, this is not a philosophy of science that can be done in isolation. It demands genuine engagement with the work and thinking of scientists. But the aim is not simply to catalog what scientists do. The aim is to analyze and refine the practical tools of scientific judgment, however rough or messy, in order to contribute practical, methodological insights about the state of research, the diversity of heuristics, the sources of error, and the promise of novel hypotheses.

Bibliography

Anderson, J. A. (2006). "The Ethics and Science of Placebo-Controlled Trials: Assay Sensitivity and the Duhem-Quine Thesis", *Journal of Medicine and Philosophy*, Vol. 31, pp.65-81.

Angrist, J.D., G. W. Imbens, D. B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables", *Journal of the American Statistical Association*, Vol. 91, pp.444-470.

Ariew, R. (1984). "The Duhem Thesis", *The British Journal for the Philosophy of Science*, Vol. 35.

Batterman, R. (2009). "Idealization and Modeling", *Synthese*, Vol. 169, pp.427-446.

Begg C.B., M.K. Cho, S. Eastwood, et al. (1996). "Improving the Quality of Reporting of Randomized Controlled Trials: The CONSORT Statement", *Journal of the American Medical Association*, Vol. 276, pp.637-639.

Boorman, S. A. and P. R. Levitt (1973). "Group Selection on the Boundary of a Stable Population", *Theoretical Population Biology*, Vol. 4, pp.85-128.

Boros, L., C. Chuange, F. Butler, and J. Bennett (1985). "Leukemia in Rochester: A 17-year-experience with an Analysis of the Role of Cooperative Group (ECOG) Participation", *Cancer*, Vol. 56, pp.2161-2169.

Brasher, P., S. Shapiro and D. Sackett (2004). "Final Report of the National Placebo Working Committee on the Appropriate Use of Placebos in Clinical Trials in Canada", in *National Placebo Initiative*, Canada Institutes of Health Research.

Burman W., S. Goldberg, J. Johnson, G. Muzanye, M. Engle, A. Mosher, S. Choudhri, C. Daley, S. Munsiff, Z. Zhao, et al. (2006). "Moxifloxacin Versus Ethambutol in the First 2 months of Treatment for Pulmonary Tuberculosis", *American Journal of Respiratory and Critical Care Medicine*, Vol. 174, pp.331-338.

Charnov, E. L. and J. R. Krebs (1975). "The Evolution of Alarm Calls: Altruism or Manipulation?", *American Naturalist*, Vol. 109, pp.107-112.

Cohen, D. and I. Eshel (1976). "On the Founder Effect and the Evolution of Altruistic Traits", *Theoretical Population Biology*, Vol. 10, pp.276-302.

Conde M., A. Efron, C. Loredo, G. De Souza, N. Gracxa, M. Cezar, M. Ram, M. Chaudhary, W. Bishai, A. Kritski, and R. Chaisson (2009). "Moxifloxacin Versus Ethambutol in the Initial Treatment of Tuberculosis: A Double-blind, Randomised, Controlled Phase II Trial", *The Lancet* Vol. 53, pp.1314-1319.

Culp, S. (1994). "Defending Robustness: The Bacterial Mesosome as a Test Case", *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Assoication*, Vol. 1, pp.46-57.

Curson, D. A. et al. (1986). "Does Short Term Placebo Treatment of Chronic Schizophrenia Produce Long Term Harm?" *British Medical Journal (Clinical Research Edition)*, Vol. 293: pp. 726-728.

D'Arms, J., R. Batterman, and K. Gorny (1998). "Game Theoretic Explanations and the Evolution of Justice", *Philosophy of Science*, Vol. 65, pp.76-102.

Darling, K. (2002). "The Complete Duhemian Underdetermination Argument: Scientific Language and Practice", *Studies in History and Philosophy of Science*, Vol. 33, pp.511-533.

Dickenson, J., V. Aber, and D. Mitchison (1977). "Bactericidal Activity of Streptomycin, Isoniazid, Rifampin, Ethambutol, and Pyrazinamide Alone and in Combination Against Mycobacterium Tuberculosis", *American Review of Respiratory Disease*, Vol. 116, pp.627-635.

Dorling, J. (1979). "Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem", *Studies in History and Philosophy of Science*, Vol. 10.

Dorman S., J. Johnson, S. Goldberg, et al. (2009). "Substitution of Moxifloxacin for Isoniazid During Intensive Phase Treatment of Pulmonary Tuberculosis", *American Journal of Respiratory and Critical Care Medicine*, Vol. 180, pp.273-280.

Duhem, P. (1906). *La Théorie Physique, son objet et sa structure*, Chevalier et Rivière. Translated as Duhem 1991.

Duhem, P. (1915). *La Science Allemande*, A. Hermann et Fils. Translated as Duhem 1991b.

Duhem, P. (1991a). *The Aim and Structure of Physical Theory*, Princeton University Press.

Duhem, P. (1991b). *German Science*, Open Court Publishing.

Eshel, I. (1972). "On the Neighbor Effect and the Evolution of Altruistic Traits", *Theoretical Population Biology*, Vol. 3, pp.258-277.

Fedson, D. (2005). "Preparing for Pandemic Vaccination: An International Policy Agenda for Vaccine Development", *Journal of Public Health Policy*, Vol. 26, pp.4-29.

Fenner, F. and F. N. Ratcliff (1965), *Myxomatosis*. London: Cambridge University Press.

Fergusson, D., K. C. Glass, D. Waring, S. Shapiro (2009). “Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials”, *British Medical Journal*, Vol. 328, pp. 432-434.

Fowler, A. C. (1997). *Mathematical Models in the Applied Sciences* Cambridge Texts in Applied Mathematics, Cambridge University Press.

Freedman, B. (1987). “Equipoise and the Ethics of Clinical Research”, *The New England Journal of Medicine*, Vol. 317, pp.141-145.

Freedman, B. (1990). “Placebo-Controlled Trials and the Logic of Clinical Purpose”, *The Hastings Center Report*, Vol. 12, pp.1-6.

Freedman, B., K. C. Glass, C. Weijer (1996). “Placebo Orthodoxy in Clinical Research II: Ethical, Legal, and Regulatory Myths”, *The Journal of Law, Medicine, and Ethics*, Vol. 24: pp. 252-259.

Friedman, L. M., C. D. Furberg, D. L. DeMets (1998). *Fundamentals of Clinical Trials*, 3rd edition. Springer.

Fuks A., C. Weijer, B. Freedman, S. Shapiro, M. Skrutkowska, and A. Riaz (1998). “A study in Contrasts: Eligibility Criteria in a Twenty-Year Sample of NSABP and POG Clinical Trials”, *Journal of Clinical Epidemiology*, Vol. 51, pp.69-79.

Gadgil, M. (1975). “Evolution of Social Behaviors Through Interpopulation Selection”, *Proceedings of the National Academy of Sciences, U.S.A.*, Vol. 72, pp.1199-1201.

Gelfand, L. A., D. R. Strunk, X. M. Tu, R. E. S. Noble, and R. J. DeRubeis (2006). “Bias Resulting from the Use of ‘Assay Sensitivity’ as an Inclusion Criterion for Meta-Analysis”, *Statistics in Medicine*, Vol. 25, pp.943-955.

Gigerenzer G. and W. Gaissmaier (2011). “Heuristic Decision Making”, *Annual Review of Psychology*, Vol. 62, pp.451-482.

Gillespie S. and O. Billington (1999). “Activity of Moxifloxacin Against Mycobacteria”, *Journal of Antimicrobial Chemotherapy*, Vol. 44, pp.393-395.

Gillespie S., R. Gosling, L. Uiso, N. Sam, E. Kanduma, and T. McHugh (2005). “Early Bactericidal Activity of a Moxifloxacin and Isoniazid Combination in Smear-positive Pulmonary Tuberculosis”, *Journal of Antimicrobial Chemotherapy*, Vol 56, pp.1169-1171.

Gilpin, M. E. (1975). *Group Selection in Predator-Prey Communities*, Princeton: Princeton University Press.

Gosling R., L. Uiso, N. Sam, E. Bongard, E. Kanduma, M. Nyindo, R. Morris, and S. Gillespie (2003). “The Bactericidal Activity of Moxifloxacin in Patients with Pulmonary Tuberculosis”, *American Journal of Respiratory and Critical Care Medicine*, Vol. 168, pp.1342-1345.

Green L. and D. Mehr (1997). "What Alters Physicians' Decisions to Admit to the Coronary Care Unit?", *Journal of Family Practice*, Vol. 45, pp.219226.

Griesemer, J. R. and M. J. Wade (1988). "Laboratory Models, Causal Explanation, and Group Selection", *Biology and Philosophy*, Vol. 3, pp.67-96.

Gunderson, L. and C. Holling (2002). *Panarchy: Understanding Transformations in Human and Natural Systems*, Island Press.

Hacking, I. (1983). *Representing and Intervening*, Cambridge University Press.

Hjorth, M., E. Holmberg, S. Rodjer, and J. Westin (1992). "Impact of Native and Passive Exlusions on the Results of a Clinical Trial in Multiple Myeloma", *British Journal of Haematology*, Vol. 80, pp.55-61.

Holling, C. (1973). "Resilience and stability of ecological systems", *Annual Review of Ecology and Systematics*, Vol. 4, pp.1-23.

Howick, J. (2009). "Questioning the Methodologic Superiority of 'Placebo' Over 'Active' Controlled Trials", *The American Journal of Bioethics*, Vol. 9: pp. 34-48.

Howson, C. and P. Urbach (1993). *Scientific Reasoning: The Bayesian Approach*, second edition, Open Court.

Hrobjartsson, A. and P. C. Gotzsche (2001). "Is the Placebo Powerless? An Analysis of Clinical Trials Comparing Placebo with No Treatment", *New England Journal of Medicine*, Vol. 259: pp. 91-100.

Hwang, I. K. and T. Morikawa (1999). "Design Issues in Noninferiority/Equivalence Trials", *Drug Information Journal*, Vol. 33: pp. 1205-1218.

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (1996). *Guideline for Good Clinical Practice E6(R1).*

International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (2000). *Choice of Control Group and Related Issues in Clinical Trials E10.*

Jen, E. (2005). "Stable or Robust? What's the Difference?" in *Robust Design: A Repertoire of Biological, Ecological, and Engineering Case Studies*, Oxford University Press.

Ji B., N. Lounis, C. Maslo, C. Truffot-Pernot, P. Bonnafous, and J. Grosset (1998). "In Vitro and In Vivo Activities of Moxifloxacin and Clinafloxacin Against Mycobacterium Tuberculosis", *Antimicrobial Agents and Chemotherapy*, Vol. 42, pp.2066-2069.

Johnson J., D. Hadad, W. Boom, C. Daley, C. Peloquin, K. Eisenach, D. Jankus, S. Debanne, E. Charlebois, E. Maciel, et al. (2006). "Early and Extended Early Bactericidal Activity of Levofloxacin, Gatifloxacin and Moxifloxacin in Pulmonary Tuberculosis", *International Journal of Tuberculosis and Lung Disease*, Vol. 10, pp.605-612.

Karjalaninen, S. and I. Palva (1989). “Do Treatment Protocols Improve End Results? A Study of the Survival of Patients with Multiple Myeloma in Finland”, *British Medical Journal*, Vol. 299, pp.1069-1072.

Kimmelman, J., C. Weijer, E. M. Meslin (2009). “Helsinki Discords: FDA, Ethics, and International Drug Trials”, *The Lancet*, Vol. 373: pp. 13-14.

Kirsch, I., B. J. Deacon, T. B. Huedo-Medina, A. Scoboria, T. J. Moore, B. T. Johnson (2008). “Initial severity and antidepressant benefits: A meta-analysis of data submitted to the food and drug administration”, *PlOS Medicine*, Vol. 5: pp. 260-268.

Kirschner, N., S. G. Pauker, J. W. Stubbs (2008). “Information on Cost-Effectiveness: An Essential Product of a National Comparative Effectiveness Program”, *Annals of Internal Medicine*, Vol. 148: pp. 956-961.

Kitcher, P., K. Sterelny, and C. K. Watters (1990). “The Illusory Riches of Sober’s Monism”, *Journal of Philosophy*, Vol. 87, pp.158-161.

Lakatos, I. (1965). “Falsification and the Methodology of Scientific Research Programmes”, in *Criticism and the Growth of Knowledge: proceedings of the international colloquium in the philosophy of science*, Vol. 4, ed. I. Lakatos and A. Musgrave, Cambridge University Press.

Laudan, L. (1965). “Grünbaum on ‘the Duhemian argument”’, *Philosophy of Science*, Vol. 32:3, pp.295-299.

Levin, B. and W. Kilmer (1974). “Interdemic Selection and the Evolution of Altruism”, *Evolution*, Vol. 28, pp.527-545.

Levins, R. (1966). “The Strategy of Model Building in Population Biology”, *American Scientist*, Vol. 54, No. 4, pp.421-431.

Levins, R. (1970). “Extinction”, in “Some Mathematical Questions in Biology: Lectures on Mathematics in the Life Sciences”, *American Mathematical Society*, Vol. 2, pp.75-108.

Levins, R. (1993). “A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science”, *The Quarterly Review of Biology*, Vol. 68, pp.547-555.

Lewontin, R. C. (1970). “The Units of Selection”, *Annual Review of Ecology and Systematics*, Vol. 1, pp.1-18.

Lloyd, E. A. (2005). “Why the Gene Will Not Return”, *Philosophy of Science*, Vol. 72, pp.287-310.

Lounis N., A. Bentoucha, C. Truffot-Pernot, B. Ji, R. O’Brien, A. Vernon, G. Roscigno, J. Grosset (2001). “Effectiveness of Once-weekly Rifapentine and Moxifloxacin Regimens Against Mycobacterium Tuberculosis in Mice”, *Antimicrobial Agents and Chemotherapy*, Vol. 45, pp.3482-3486.

Lyapunov A. (1966). *Stability of Motion*, Academic Press.

MacKenzie, W., C. Heilig, L. Bozeman, J. Johnson, G. Muzanye, D. Dunbar, K. Jost Jr., L. Diem, B. Metchock, K. Eisenach, S. Dorman, and S. Goldberg (2011). "Geographic Differences in Time to Culture Conversion in Liquid Media: Tuberculosis Trials Consortium Study 28. Culture Conversion Is Delayed in Africa", *PLoS One*, Vol. 6, e18358.

Matessi, C. and S. Jayakar (1973). "A Model for the Evolution of Altruistic Behavior", *Genetics*, Vol. 74, §174.

Matessi, C. and S. D. Jayakar (1976). "Conditions for the Evolution of Altruism Under Darwinian Selection", *Theoretical Population Biology*, Vol. 9, pp.360-387.

Maynard Smith, J. (1964). "Group Selection and Kin Selection", *Nature*, Vol. 201, pp.1145-1147.

Maynard Smith, J. (1967). "Group Selection", *The Quarterly Review of Biology*, Vol. 51, pp.277-283.

Mitchell, S. (2000). "Dimensions of Scientific Law", *Philosophy of Science*, Vol. 67, pp.242-265.

Miyazaki E., M. Miyazaki, J. Chen, R. Chaisson, and W. Bishai (1999). "Moxifloxacin (BAY12-8039), a New 8-methoxyquinolone, is Active in a Mouse Model of Tuberculosis", *Antimicrobrial Agents and Chemotherapy*, Vol. 43, pp.85-89.

Moher, D., K. F. Schulz, D. G. Altman (2001). "The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomised Trials", *The Lancet*, Vol. 357: pp.1191-1194.

Muldoon, R. (2007). "Robust Simulations", *Philosophy of Science*, Vol. 75, pp.873-883.

Needham, P. (2000). "Duhem and Quine", *Dialectica*, Vol. 54, pp.109-132.

Nuermberger E., T. Yoshimatsu, S. Tyagi, R. O'Brien, A. Vernon, R. Chaisson, W. Bishai, and J. Grosset (2004a). "Moxifloxacin-containing Regimen Greatly Reduces Time to Culture Conversion in Murine Tuberculosis", *American Journal of Respiratory and Critical Care Medicine*, Vol. 169, pp.421-426.

Nuermberger E., T. Yoshimatsu, S. Tyagi, K. Williams, I. Rosenthal, R. O'Brien, A. Vernon, R. Chaisson, W. Bishai, and J. Grosset (2004b). "Moxifloxacin-containing Regimens of Reduced Duration Produce a Stable Cure in Murine Tuberculosis", *American Journal of Respiratory and Critical Care Medicine*, Vol. 170, pp.1131-1134.

Orzack, S. H. and E. Sober (1993). "A Critical Assessment of Levins's *The Strategy of Model Building in Population Biology* (1966)", *The Quarterly Review of Biology*, Vol. 68, pp.533-546.

Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*, Addison-Wesley.

Perel, P., I. Roberts, E. Sena, P. Wheble, C. Briscoe, P. Sandercock, M. Macleod, L. Mignini, P. Jayaram, and K. Khan (2007). "Comparison of Treatment Effects between Animal Experiments and Clinical Trials: Systematic Review", *British Medical Journal*, Vol. 334, pp. 197-200.

Pletz M., A. De Roux, A. Roth, K. Neumann, H. Mauch, and H. Lode (2004). "Early Bactericidal Activity of Moxifloxacin in Treatment of Pulmonary Tuberculosis: A Prospective, Randomized Study", *Antimicrobial Agents and Chemotherapy*, Vol. 48, pp.780-782.

Rasmussen, N. (1993). "Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope", Studies in the History and Philosophy of Science, Vol. 24, pp.227-265.

Rustomjee R, C. Lienhardt, T. Kanyok and Gatifloxacin for TB (OFLOTUB) study team (2008). "A Phase II Study of the Sterilising Activities of Ofloxacin, Gatifloxacin and Moxifloxacin in Pulmonary Tuberculosis", *International Journal of Tuberculosis and Lung Disease*, Vol. 12, pp.128-138.

Sampson, H., C. Weijer, and D. Pullman (2009). "Research Governance Lessons from the National Placebo Initiative", *Health Law Review*, Vol. 17: pp. 26-32.

Schwartz, D. and J. Lellouch (1967). "Explanatory and Pragmatic Attitudes in Therapeutical Trials", *Journal of Chronic Disease*, Vol. 20, pp.637-648.

Shapiro, A. K. (1970). "Placebo Effects in Psychotherapy and Psychoanalysis", *Journal of Clinical Pharamcology*, Vol. 10: pp.73-78.

Simon, H. (1957). "A Behavioral Model of Rational Choice", in H. A. Simon *Models of Man*, Wiley.

Smale, S. (1980). "What is Global Analysis?", in *The Mathematics of Time: Essays on Dynamical Systems, Economic Processes, and Related Topics*, Springer-Verlag, pp. 84-89.

Spiro, H. M. (1986). *Doctors, Patients, and Placebos*, Yale University Press.

Stass, H., A Dalhoff, D. Kubitza, and U. Schühly (1998). "Pharmacokinetics, Safety, and Tolerability of Ascending Single Doses of Moxifloxacin, a New 8-Methoxy Quinolone, Administered to Healthy Subjects", *Antimicrobial Agents and Chemotherapy*, Vol. 42, pp.2060-2065.

Steenwinkel, J., G. Knegt, M. Kate, A. Belkum, H. Verbrugh, K. Kremer, D. Soolingen, and I. Bakker-Woudenberg (2010). "Timekill Kinetics of Anti-tuberculosis Drugs, and Emergence of Resistance, in Relation to Metabolic Activity of Mycobacterium Tuberculosis", *Journal of Antimicrobial Chemotherapy*, Vol. 65, pp.2582-2589.

Stegenga, J. (2009). "Robustness, Discordance, and Relevance", *Philosophy of Science*, Vol. 76, pp.650-661.

Sterelny, K. and P. Kitcher (1988). “The Return of the Gene”, *Journal of Philosophy*, Vol. 85, pp.339-361.

Temple R. (1983). “Difficulties in Evaluating Positive Control Trials”, *Proceedings of the American Statistical Association*, pp.1-7.

Temple R. and S. S. Ellenberg (2000). “Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments, Part 1: Ethical and Scientific Issues”, *Annals of Internal Medicine*, Vol. 133: pp. 455-463.

Temple R. and S. S. Ellenberg (2000). “Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments, Part 2: Practical Issues and Specific Cases”, *Annals of Internal Medicine*, Vol. 133: pp. 464-470.

Thorpe, K., M. Zwarenstein, A. Oxman, S. Treweek, C. Furberg, D. Altman, S. Tunis, E. Bergel, I. Harvey, D. Magid, and K. Chalkidou (2009). “A Pragmatice-explanatory Continuum Indicator Summary (PRECIS): A Tool to Help Trial Designers”, *Journal of Clinical Epidemiology*, Vol. 62, pp.464-475.

Todhunter, I. (1876). *William Whewell, D.D.: an account of his writings, with selections from his literary and scientific correspondence*, Vol. 2, Macmillan and Co.

Tversky A. and D. Kahneman (1976). “Judgment under Uncertainty: Heuristics and Biases”, *Science*, Vol. 185:4157, pp.1124-1131.

United States Code of Federal Regulations (2009). Title 21, Volume 7, §601.2.

United States Food and Drug Administration (2010). “Centers and Offices”, U.S. Department of Health and Human Services, retrieved March 2010. URL = http://www.fda.gov/AboutFDA/CentersOffices/default.htm

United States Food and Drug Administration (2010). “FDA Fundamentals”, U.S. Department of Health and Human Services, retrieved March 2010. URL = http://www.fda.gov/AboutFDA/Basics/ucm192695.htm

Van der Worp, H., D. Howells, E. Sena, M. Porritt, S. Rewell, V. OCollins2, M. Macleod (2010). “Can Animal Models of Disease Reliably Inform Human Studies?”, *PLoS Medicine*, Vol. 7, pp.1-8.

Wade, M. J. (1976), “Group Selection Among Laboratory Populations of *Tribolium*”, *Proceedings of the National Academy of Sciences, U.S.A.*, Vol. 73, pp.4606-4607.

Wade, M. J. (1977). “An experimental study of group selection”, *Evolution*, Vol. 31, pp.134-153.

Wade, M. J. (1978). “A Critical Review of the Models of Group Selection”. *The Quarterly Review of Biology*, Vol. 53, No. 2, pp.101-114.

Walker, B., L. Gunderson, A. Kinzig, C. Folke, S. Carpenter, and L. Schultz (2006). "A Handful of Heuristics and Some Propositions for Understanding Resilience in Social-Ecological Systems", *Ecology and Society*, Vol. 11.

Walker, B. and D. Salt (2006). *Resilience Thinking*, Island Press.

Wang J. Y., J. T. Wang, T. Tsai, C. Hsu, C. Yu, P. Hsueh, L. Lee, and P. Yang (2009). "Adding Moxifloxacin is Associated with a Shorter Time to Culture Conversion in Pulmonary Tuberculosis", *International Journal of Tuberculosis and Lung Disease*, Vol. 14, pp.65-71.

Weijer, C., B. Freedman, A. Fuks, J. Robbins, S. Shapiro, and M. Skrutkowska (1996). "What Difference Does It Make to Be Treated on a Clinical Trial? A Pilot Study", *Clinical and Investigative Medicine*, Vol. 19, pp.179-183.

Weisberg, M. (2006). "Robustness Analysis", *Philosophy of Science*, Vol. 73, pp.730-742.

Weisberg, M. and K. Reisman (2008). "The Robust Volterra Principle", *Philosophy of Science*, Vol. 75, pp.106-131.

Whewell, W. (1840). *The Philosophy of the Inductive Sciences*, Parker and Deighton.

Williams, G. (1966). *Adaptation and Natural Selection*, Princeton University Press.

Wilson, D. S. (1975). "A General Theory of Group Selection", *Proceedings of the National Academy of Sciences, U.S.A.*, Vol. 72, pp.143-146.

Wilson, D. S. (1977). "Structured Demes and the Evolution of Group Advantageous Traits", *American Naturalist*, Vol. 111, pp.157-185.

Wilson, D. S. (1983). "The Group Selection Controversy: History and Current Status", *Annual Review of Ecology and Systematics*, Vo. 14, pp.159-187.

Wilson, D. S. (1989). "Levels of Selection: An Alternative to Individualism in Biology and the Human Sciences", *Social Networks*, Vol. 11, pp.257-272.

Wilson, M. (2007). *Wandering Significance*, Oxford University Press.

Wilson, R. (2004). "Test Cases, Resolvability, and Group Selection: A Critical Examination of the Myxoma Case", *Philosophy of Science*, Vol. 71, pp.380-401.

Wimsatt, W. (1981). "Robustness, Reliability, and Overdetermination", in M. Brewer and B. Collins (eds.), *Scientific Inquiry and the Social Sciences*, San Francisco: Jossey- Bass, pp.124-163.

Wimsatt, W. (1982). "Reductionist Research Strategies and Their Biases in the Units of Selection Controversy", in *Conceptual Issues in Ecology*, ed. E. Saarinen, D. Reidel Publishing.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*, Harvard University Press.

World Medical Association (2008). "Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects".

World Medical Organization (1964). "Declaration of Helsinki: Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects", printed in *British Medical Journal* (1996), Vol. 313, pp.1448-1449.

Wright, S. (1945). "*Tempo and Mode in Evolution*: A Critical Review", *Ecology*, Vol. 26, pp.415-419.

Wynne-Edwards, V. (1962). *Animal Dispersion in Relation to Social Behavior*, Oliver and Boyd.

Zhu, M., W. Burman, J. Starke, J. Stambaugh, P. Steiner, A. Bulpitt, D. Ashkin, B. Auclair, S. Berning, R. Jelliffe, G. Jaresko, C. Peloquin (2004). "Pharmacokinetics of ethambutol in children and adults with tuberculosis", *International Journal of Tuberculosis and Lung Disease*, Vol. 8, pp.1360-1367.

Curriculum Vitae

Name:	Spencer Hey
Post-Secondary Education and Degrees:	University of Western Ontario 2006 - 2011 Ph.D. University of Western Ontario 2005 - 2006 M.A. University of Illinois at Chicago 2002 - 2004 B.A.
Honours and Awards:	Canadian Bioethics Society Student Abstract Award 2009 Rotman Institute Graduate Research Award 2007 - 2010 University of Western Ontario Graduate Research Scholarship 2005 - 2010
Related Work Experience:	Lecturer The University of Western Ontario 2010 - 2011 Teaching Assistant The University of Western Ontario 2005 - 2009

Publications:

Spencer Phillips Hey and Charles Weijer (forthcoming). "What Questions Can a Placebo Help Answer?" In *Proceedings of the McGill Placebo Workshop* (under consideration at Oxford University Press).

Presentations:

"Crossing Disciplines: From TB Trials to Philosophy," Tuberculosis Trials Consortium 30th Semi-Annual Group Meeting, October 2011

"Lexicon of Invariance," International Society for the History, Philosophy, and Social Studies of Biology Biannual Meeting, University of Utah, July 2011

"What Questions Can a Placebo Help Answer?" Co-presented with Charles Weijer, Placebo Workshop, McGill University, July 2010

"Robustness and the Burden of Clinical Equipoise," Objectivity in Science Conference, University of British Columbia, June 2010

"The Assay Sensitivity Problem," co-presented with Charles Weijer, Canadian Philosophical Association Annual Meeting, Concordia University, June 2010

"Robustness and Group Selection," Canadian Society for the History and Philosophy of Science, Concordia University, May 2010

"Equipoise, Experimental Design, and Inferential Strength," Canadian Bioethics Society Annual Meeting, May 2009

"The Duhem Problem and Duhem's Solution," 24*th* Annual Boulder Conference on the History and Philosophy of Science, October 2008