On the Robustness of Economic Models

Johanna Thoma

Faculty of Philosophy at Erasmus University Rotterdam, and Erasmus Institute for Philosophy and Economics

Research Master Philosophy and Economics: Thesis

Title: "On the Robustness of Economic Models"

Author: Johanna Marie Thoma, BA (Hons.)

Student Number: 347840

Supervisor: Conrad Heilmann

Advisor: Marcel Boumans

ECTS: 30

Word Count: 43,031

Date of Completion: 30.05.2012

On the Robustness of Economic Models

Contents:

1. Introduction

"[O]ur truth is the intersection of independent lies."
(Levins 1966, p.423)

Theoretical economists spend much of their time deriving the same results using models that make slightly different assumptions. Kuorikoski et al. (2010) have recently suggested that this practice may provide us with a solution to one of the most notorious problems in the philosophy of economics: the problem of how we can learn from the highly idealised models of theoretical economics. Following a tradition originating in the philosophy of biology and the work of Levins (1966, 1968), they suggest that by showing that a modelling result stays the same when some assumptions are varied, we gain confidence in inferences from the model, since we learn that some problematic or arbitrary modelling assumptions were irrelevant to a modelling result of interest. Thereby the practice of deriving the same result with models making various different assumptions, which we will call *robustness analysis*, is said to offer a kind of confirmation.

This thesis elaborates two distinct arguments that have been made for the confirmatory value of robustness analysis, and ultimately rejects them. It will show that, for both arguments, a robustness analysis which could potentially reap these epistemic benefits is much more demanding than is commonly acknowledged. And indeed, a case study on models of herd behaviour will show that the economists working in this field do not do enough for robustness analysis to serve a confirmatory function, and that in fact it is hard to imagine that robustness analysis could potentially be confirmatory.

What the main case study of this thesis also shows, however, is that economists are not in fact after confirmation when they accumulate alternative models that share a core result with an original model. So while the philosophical accounts of the value of robustness analysis that will be discussed in this thesis fail, this does not need to mean that the practice of robustness analysis is pointless. The last chapter of this thesis will explore some alternative interpretations of the purpose of robustness analysis.

1.1. Motivation and Background

This thesis connects with and responds to three pressing problems in the philosophy of economics. The first is the unresolved problem of how we can learn from the simple, highly idealised models of theoretical economics about real world targets that differ in a variety of ways from these models. The second is the puzzling practice of robustness analysis, which occurs when economists repeatedly derive the same results using models that make different

assumptions. This practice is widespread in theoretical economics, and evidently thought of as useful. Yet economists rarely make explicit why it is such a good thing when a result has been found to be robust. Making sense of this practice is hence another unresolved problem. The third is to respond to a philosophical position recently expressed by Kuorikoski et al. (2010), but which goes back to Levins (1966), Wimsatt (1987) and Weisberg (2006a, 2006b). This position brings together the two open questions just mentioned in that it argues that robustness analysis is instrumental in learning from the highly idealised models of theoretical economics: Robustness analysis can increase our confidence in inferences from our models.

1.1.1. Learning from Models

To start with the first problem, the question of how we can learn from economic models is one of the most long-standing problems in the philosophy of economics. The mathematical models in question are typically simple, analytically solvable, and, taken as descriptions of the real world, they are incomplete and literally false in many respects. Following Friedman (1953), the debate has been framed around the question of how to justify the "unrealistic assumptions" made in these models; more recently, Sugden (2000) claimed the central problem lies in establishing in virtue of what we can make inferences from an artificial model world to the real world.

What makes the issue very pressing for economic methodologists is that many of the traditional accounts of how we can learn from scientific models *in general* cannot be easily applied to the case of models in economics. One traditional view, defended by Cartwright (1999), is that idealised economic models can help us identify capacities - causal tendencies in the world that make stable differences across a variety of situations. Models help us study capacities by isolating them from disturbances. However, as Cartwright (2006) herself later argues, economic models do not generally identify capacities. Another standard approach is that the real world needs to instantiate features of the model in order for us to learn about the world from the model. One proposal (see McMullin 1985) is to say that the world needs to instantiate those features of the model which cannot be 'de-idealised', i.e. made more realistic without changing the outcome of a model. As Reiss (2007) points out, the problem with this account for economics is that economic models are usually too sensitive to changing assumptions to allow for enough de-idealisation to make a model apply to a phenomenon of interest on this account.

Despite these problems of justifying the kind of idealisations economics makes, very simple and highly idealised models are still ubiquitous in economics, and have not been replaced, for instance, by more complex simulations which may contain less stark idealisations. In order to account for this, some less demanding accounts of learning from models have been developed. For instance, Alexandrova (2008) sees models as open formulae which provide us with hypotheses for experimental testing. This accords models a fairly minimal role in aid of more empirical research. Sugden (2000, 2009) thinks of models as "credible worlds" which can tell us something about the real world in virtue of economists' plausibility judgements. The position Kuorikoski et al. defend can be seen in this tradition of looking for a less demanding account that can at least tell us how we can gain confidence in our inferences from models, using a

method that is in fact widespread in economic modelling practice: they argue that when we find a modelling result to be robust, this result has additional epistemic credence.

1.1.2. Robustness Analysis

The second problem relates to the practice of robustness analysis in economic modelling. Robustness may refer to a variety of things, as chapter 3 explores, but in general, the idea is that some result is stable when various ways of determining that result are applied: for instance when different means of measuring a quantity lead to the same result. Robustness has been considered a fruitful concept in the philosophy of science in general, playing an important role in confirmation. The idea that when a number of independent sources of evidence support the same hypothesis, our confidence in it should increase, is wide-spread. It is elaborated and generalised, for instance, by Wimsatt (1981). Wimsatt calls robustness analysis "triangulation via independent ways of determination", and grounds its power in the idea that it would be a remarkable coincidence if independent ways of determining a result pointing in the same direction did not point in the right direction.

Robustness is also an important concern for economists. Econometricians are very much concerned to show that the results they derive from data are stable when an econometric model is changed in various respects. The claim that an econometric result is 'fragile' is usually taken to be a severe criticism. There is also no lack of philosophical literature dealing with robustness and fragility in econometrics, with a recent symposium in the *Journal for Economic Methodology* (2006: 13 (2)) devoted to the subject. Our main concern, however, is not econometrics, but robustness in theoretical modelling: In theoretical economic modelling, as we will see in chapter 2, there is much talk of robustness, too. What is usually meant here is that after presenting a model, economists offer modifications to their original model which still preserve the main result they were interested in. Sometimes the term 'robustness' is also used when economists offer alternative models that yield the same result as a model published by a different economist earlier. Theoretical economists take the robustness of a modelling result to be a good thing, but, much like econometricians, are rarely explicit about why they think so.

1.1.3. Arguments for the Confirmatory Value of Robustness Analysis

Kuorikoski at al. (2010) provide an answer to the question of what robustness analysis is good for by seeing it as a response to the problem of learning from idealised models. They take Wimsatt's ideas on the confirmatory value of robustness and apply it to the case of robustness analysis in theoretical modelling. In line with what we just said about the problems of learning from economic models, they acknowledge that when we try to make inferences from models we are usually worried about a number of 'unrealistic' assumptions. But by triangulating a result with models using alternative sets of assumptions, we can lose worries about the unrealistic assumptions of the individual models – we gain confidence that the result we take from the model did not depend on arbitrary modelling assumptions. This procedure for gaining confidence in inferences from models, they claim, is central to economic modelling practice, since they maintain that theoretical economic modelling essentially *is* robustness analysis: In a collective process, economists refine and adapt models, or use modified models to criticise other people´s models.

The main part of this thesis will be concerned with exploring and rejecting the idea that robustness analysis can give us a kind of confirmation and thereby help with learning from theoretical models. We will show that Kuorikoski et al.'s argument is ambiguous between two ways of arguing for the claim that robustness analysis is confirmatory. But neither argument succeeds, because economic practice does not fulfil the requirements necessary for either argument, which I will show using a case study on models of herd behaviour. So Kuorikoski et al. fail to make sense of the economic practice they set out to justify. And indeed, it seems implausible that it could ever do so, even if we were to change economic practice.

Instead, the last part of the thesis suggests that economists are not even trying to get at confirmation when they conduct robustness analysis. Instead, having a variety of models that yield the same result is seen as beneficial for a number of other reasons. In particular, having this variety of models can be seen as being useful for the purpose of explanation. It may also allow for the inference that a phenomenon of interest occurs in a variety of situations in the target, just in case we think each of our models allows for reliable inferences in the first place. What robustness does not do, cannot do, and is not intended to do is increase our confidence in individual inferences from models.

1.2. Chapter Summaries

1.2.1. Part I

Part I will offer an introduction to robustness and provide the necessary background for the main argument of the thesis. Chapter 2 introduces the economic practice of robustness analysis, taking Banerjee's model of herd behaviour as an example. Robustness analysis is the practice of economists deriving the same result from theoretical models making different assumptions. In economics, robustness is typically seen as a good thing, but it is rarely explained what exactly its virtue is. This is a puzzle that the accounts studied in later chapters attempt to provide an answer to.

Chapter 3 provides an overview of different kinds of robustness in the philosophical and wider scientific literature, and is roughly based on Woodward's (2006) taxonomy of four kinds of robustness. Most importantly, this chapter distinguishes between robustness in the target (stability of some property or causal relationship within the real world system we are studying) and methodological robustness (stability of the result from a scientific investigation). It also distinguishes between two ways of arguing for the confirmatory value of robustness: robustness as the agreement of a variety of evidence on the one hand, and inferential robustness on the other. It is important to make the latter distinction because the success of each of these kinds of robustness in making inferences more reliable depends on different factors: Independence of sources of evidence in the first case, and exhaustiveness of options tried in the second. Both distinctions just mentioned will turn out to be important for the main argument of the thesis. Further, the last type of robustness in Woodward (2006), namely derivational robustness, will be shown to be underdeveloped. Derivational robustness refers to the robustness of a result under varying assumptions in theoretical modelling, which is

precisely what the practice identified in chapter 2 aims at. We are hence left with the same puzzle as in chapter 2 – what is the use of robustness of results in theoretical modelling?

1.2.2. Part II

Part II deals with one proposal of what the virtue of robustness analysis is: the alleged confirmatory value of robustness analysis. It introduces the views of the most prominent philosophers of science who have argued for the confirmatory value of robustness analysis, but ultimately rejects them. The authors defending these accounts have seen robustness analysis as a way to make more reliable inferences from models and hence have seen robustness analysis as part of an answer to the debate on how we learn from simple economic that we have just introduced, and that will be elaborated in chapter 4.

Chapter 4 introduces approaches to robustness in theoretical modelling that see it as confirmatory. It argues that there are two distinct ideas that essentially see robustness in theoretical modelling as either a species of agreement of a variety of evidence or of inferential robustness, and accordingly see robustness analysis as confirmatory. This is remarkable because Woodward had suggested that robustness in theoretical modelling is something altogether different from these kinds of robustness: This suggests that Woodward's derivational robustness is not a distinct type of robustness after all. Distinguishing these two ideas and interpreting existing accounts along the lines of these distinct types of robustness is illuminating, because, as chapter 3 argues, the two types of robustness have different normative credentials, i.e. see confirmatory value in robustness for different reasons. Being clear on these reasons is crucial when criticising these two accounts.

Chapter 5, after giving more flesh to the case study on herd behaviour introduced in chapter 2, goes on to analyse and reject the first approach to arguing for the confirmatory value of robustness in modelling, namely seeing robustness analysis as a way to establish the agreement of a variety of evidence for a hypothesis. For this account to be successful, the different models compared in a robustness analysis cannot share similar biases, and cannot have been selected for according to whether they produce the result of interest. But in actual economic practice, as the case study demonstrates, this is not guaranteed. And in fact, this chapter makes the case that economic practice should not, and could not easily be changed to fulfil these requirements.

Chapter 6 analyses and rejects the second approach to arguing for the confirmatory value of robustness analysis. According to this approach, by varying individual assumptions in a model, we gain confidence that an assumption is not driving a result – several of the authors we will look at speak of this as the discovery of robust theorems, which are conditional statements linking only those assumptions which are relevant for the derivation of a result with a modelling result of interest. If the particular assumption we found to be unimportant had been cause for worry, for instance because we judge it to be unrealistic, then this is said to make inferences from the model more reliable. This chapter will argue that we are usually not licensed to conclude from robustness analysis that particular assumptions are unimportant for the derivation of a result. Further, even if we did find this out, this does not generally license increased confidence in inferences from the model. The main problem is that whether an

assumption is relevant, or whether its being 'unrealistic' is cause for worry about inferences from the model can depend on the rest of the model. Accounts of robustness analysis as inferential robustness rely on an untenable view of models as deriving their credibility for the purpose of inference from the degree of truthfulness and relative relevance of all its assumptions looked at in isolation. Hence both attempts to argue that robustness analysis is confirmatory fail.

1.2.3. Part III

Part III draws lessons from the failure of the arguments for the confirmatory value of robustness analysis and suggests alternative interpretations of what economists are doing when they conduct robustness analysis. It takes a closer look at prevalent features of economic practice and argues that economists were never aiming at getting confirmation out of robustness analysis in the first place. Seeing that getting confirmation out of robustness analysis seems to be such a hopeless endeavour, instead of trying to make economic practice fit the philosophical accounts presented in part II, it is worthwhile to look into what economists are in fact aiming at.

Chapter 7 suggests that when economists conduct robustness analysis, they either introduce models they simply judge to be better than an original one, and see it as a kind of de-idealisation. Or they introduce a model that is substantively different to the point that the model will apply in a different kind of situation. In neither case can we say that a collection of models that agrees on a main result combine to increase our confidence in one hypothesis or one kind of inference from a model: economists are not aiming to get confirmation out of robustness. Instead, having a collection of models that agree on a result has advantages when it comes to explanation. Further, if each of the models in a collection of models that share a result allows for reliable inferences about some part of the target system we are interested in, then the robustness in these models may teach us about robustness in the target: that a certain phenomenon is prevalent and occurs in a variety of situations in the target. In this case, methodological robustness has the purpose of teaching us about robustness in the target, not the purpose of making individual inferences more reliable.

The upshot is that robustness analysis may play an important role when it comes to explanation and learning about robustness features of a target. But this provides no answer to how to learn from idealised economic models, which was the hope of the philosophers whose arguments part II deals with.

PART I: Introducing Robustness

The idea of robustness is at once powerful and puzzling: It is frequently appealed to in philosophical and scientific argument to great effect, yet a closer analysis reveals much variety in what is meant by robustness, and what it is supposed to be doing. To provide some illustration, and to organise these diverse ideas, this part introduces both the economic practice of robustness analysis in modelling, and the general literature on robustness in the philosophy of science. This will set us up, and provide the necessary background, for part II, which brings these two themes together by discussing philosophical accounts of robustness in economic modelling.

2. A Model of Herd Behaviour and the Practice of Robustness Analysis

Much of economics proceeds in the formulation of theoretical models. The most influential models in microeconomics are typically relatively simple derivative arguments, expressed mathematically. On first inspection, they provide surprising explanations of familiar phenomena: Take Akerlof's (1970) famous model of the used cars market, which aimed to explain the sharp price differential between new cars and only slightly used cars. Before, economists had thought this price differential was to be explained by a pure preference for strictly new cars amongst the consumers. Akerlof's model suggested that the price differential could be explained by an appeal to asymmetric information, i.e. the fact that consumers cannot immediately identify the quality of a used car, whereas the car dealer knows the car's quality. Similarly influential models that offer surprising explanations are Selten's (1978) model of entry deterrence in monopolistic markets or Schelling's (1969) checkerboard model of spatial segregation.

What I want to present here is a slightly more recent model which is in many ways exemplary of how economic models are presented in microeconomic theory (see also Sugden 2000): Banerjee's (1990) model of herd behaviour. This model and other models of herd behaviour will form the main case study for this thesis. Still, I take the features that I want to highlight in the way these models have been constructed and modified to be features that can be found in theoretical modelling practice more generally. Let me first introduce Banerjee's model, which was one of the original models of herd behaviour (Bikhchandani et al. (1992) being the other one). It will serve to highlight a practice which is ubiquitous in theoretical economic modelling, namely that of deriving the same result using alternative model specifications. This is the

practice whose confirmatory value Kuorikoski et al. (2010) argue for and I want to contest, and I will refer to this practice as *robustness analysis* in the following.[1]

2.1. The Model

Banerjee´s model is designed to explain the phenomenon of what he calls 'herd behaviour': This phenomenon occurs when people do what everybody else is doing even though their private information suggests doing something quite different. Banerjee suggests a wide range of real world situations where we can witness this kind of behaviour: From restaurant choice to financial markets, and from fashion to choice of research topic in academia. A number of straightforward explanations of these phenomena come to mind, not least that people may just have a preference for doing what everybody else is doing – especially when it comes to choice of clothing or restaurant. In the case of choosing a research topic in academia, we may think that it is simply more fruitful to work on a topic that a number of other scholars work on. In the case of financial markets, we may have some notion of herd behaviour having to do something with a lack of information: Everybody is to some extent in the dark – but if somebody else makes a choice, we may think they know something that we don't.

This last idea is the thought that Banerjee develops in his model. Essentially, he asks us to see any occurrence of herd behaviour as a problem of imperfect information, where agents only have private signals about some object of choice, but try and infer something about other people´s information by observing their behaviour. The following little story conveys the intuition behind Banerjee's model: I may have some fairly reliable information that restaurant A is better than restaurant B, but if I see that restaurant B is crowded while restaurant A is virtually empty, I may think that all these people must know something about restaurant B that I don't. So I do not act on my own prior information but follow everybody else's lead. But if everybody reasons that way, then not only will we all end up in the same place, but where we end up depends on what the first few people do, even if their information was flawed, and most other people actually know better. By trying to extract valuable information from observing other people's choices, we make our own choices less responsive to our own information. But as a consequence, our choice will be less informative to others. In the end, collectively we do not use our information efficiently. Banerjee describes this as a kind of information externality.

Banerjee develops this little story into a slightly more complex, formal mathematical model. Here is a list of assumptions Banerjee explicitly makes in his formal model of herd behaviour:

[1] Kuorikoski et al. themselves use a case study from geographical economics, based on Krugman's (1991) core-periphery model of the spatial agglomeration of industry. I use my own case study because I find it richer, and it in fact makes a stronger prima facie case for the argument that Kuorikoski et al. aim to make: I have found more modifications of Banerjee's model both by the author himself and other economists than Kuorikoski et al. present in their case study on Krugman's model. But to dispense worries that my argument in this thesis unfairly relies on my selection of a case study, I will briefly discuss Kuorikoski et al.'s case study in the Appendix. This will show that most of the problems that will be identified throughout this thesis can be found here as well, or are indeed more striking.

- There are N agents with identical, risk neutral von Neumann-Morgenstern utility functions.
- There is a continuum of options i, indexed on the interval [0,1], one of which, i*, gives a fixed return of $z > 0$ while the others have a return of 0 (so all agents strictly prefer z).
- Agents receive a private signal with probability $\alpha < 1$, which tells them which option has return z. However, this signal is not perfectly reliable and indicates the truth with probability $\beta < 1$, and gives a random, uniformly distributed signal otherwise.
- Agents decide between the options sequentially (one after the other) and can observe all choices made prior to their own.
- The structure of the setup is common knowledge[2] and all agents exhibit Bayesian rationality (they maximise their expected utility and update their beliefs according to Bayes' rule).
- The following tie-breaking assumptions are made:
    - When an agent has no signal, and all others have chosen i=0, the agent chooses i=0
    - When an agent is indifferent between acting on their own signal and following another's choice, the agent follows their own signal
    - When an agent is indifferent between following the choices of a number of other agents, the agent chooses the highest i
- The model is solved for a Bayesian Nash equilibrium

The main results from the model are the following: Firstly, after the first few choices, agents will start to 'herd' on one value of i. Herding will begin with the first agent not to have a signal after the first agent to have a signal. That agent follows the highest i previously chosen (which will be the choice by the agent who has had a signal, given that all previous, signal-less agents will have chosen i=0), and all subsequent agents will follow that choice. Secondly, what value of i agents herd on is determined by the signals that the first agents happened to receive and later signals do not matter anymore for the outcome because of positive feedback from the first choices. This is a kind of excess sensitivity to early signals. The early signals may well happen to be false though, and so in the Bayesian Nash equilibrium, the probability that nobody chooses the "right" option i* is given by

$$\frac{(1-\alpha)(1-\beta)}{1-\alpha(1-\beta)}$$

which is decreasing in both α and β (the probability that an agent gets a signal, and that that signal is correct, respectively), and strictly greater 0. This result is shown to be ex ante inefficient.

As one of the original models of herd behaviour, this model has been very influential in fields as diverse as the study of markets, in particular financial markets, the study of social

[2] Each agent knows the structure of the setup, and each knows that everybody else knows the structure of the setup, and each knows that everybody else knows that everybody else knows the structure of the setup, and so on ad infinitum

phenomena, such as the behaviour of social networks, and in social learning theory. It has inspired many follow-up models in these fields, some of which we will encounter in chapter 5. The model has been so successful because it provides an explanation of why agents that receive diverse information may nevertheless quickly agree on a choice or belief – a phenomenon that appears to occur frequently in economic and other social contexts. At the same time, it explains why this kind of herding is inefficient from a social point of view: It makes obvious how information is not used efficiently.

2.2. Robustness in Banerjee's Model

Now Banerjee does not leave it at simply presenting the above model: Like many economics papers presenting models, his paper ends with a section introducing extensions and modifications to the original model. Six such extensions can be found in Banerjee's paper:

1) If the first agent to choose i* gets a higher reward than all subsequent agents, the qualitative results can be shown to remain, but herding is somewhat mitigated. Banerjee further hypothesises that in general, decreasing rewards will tend to mitigate herding, and increasing rewards will tend to increase it.
2) We can relax the informational requirements somewhat: The result remains the same when agents do not know the order of the previous choices, but only the distribution.
3) If agents can choose to wait, but waiting is costly, the results are similar.
4) If signals can be obtained by agents at a cost, there will probably be even more herding.
5) According to preliminary examinations, results are similar for a large but finite number of options.
6) The results also remain similar when there are a small number of different types of signals.

Out of these extensions, only (1) is in fact formally presented, all others are deemed to be either obvious or informed guesses by the author.

Throughout this section, Banerjee uses the word 'robust' or 'robustness'. For instance, when talking about the fourth extension, he notes that "[i]n this direction, at least, our results seem robust." (p. 816) What he apparently means is that the results of the model which he identified as the most important ones remain the same when we replace the assumption that agents receive signals randomly with the assumption that agents pay to receive a signal, and similarly for the other alterations he performs. Finding out that this is the case is evidently judged to be useful by Banerjee: He uses 'robustness' as a word with positive connotations: 'at least' we have found some robustness.

This use of the word is widespread in economic modelling: Economists try out a number of alternative specifications of their models, which preserve the results that had been identified as the crucial ones. 'Robustness' is then proclaimed and deemed to be something positive. This practice is at once familiar and puzzling. When pressed to say why robustness - in the sense that a modelling result is preserved when assumptions are varied - is a good thing there is no

immediate, obvious answer. To make sense of this practice is one of the main motivating questions of this thesis: What are economists trying to achieve when they vary assumptions and find that a result remains stable? Why is stability of a result under varying assumptions judged to be a good thing?

Kuorikoski et al. and others have thought that the value of robustness lies in its confirmatory value: we can have more confidence in inferences from models when they involve robust results, and robustness analysis is indeed the most crucial part of modelling in economics. Others have suggested that this is trying to do the impossible, namely attempting a non-empirical form of confirmation (e.g. Orzack and Sober 1993). Whether or not robustness analysis can serve as a form of confirmation will be our guiding question, and chapters 5 and 6 will argue that it cannot. Before we can tackle this question, however, we need to get some understanding of what philosophers and scientists mean when they speak of robustness, and what kinds of uses robustness is usually thought of as serving. There are a number of misunderstandings surrounding the debate about whether robustness in modelling can be confirmatory, partly stemming from the variety of ways in which the term 'robustness' is used both in the philosophical and the scientific literature. To make sense of the debate, we will hence first review the different types of robustness that can be found in the literature, and why each of them is judged to be a useful thing to have.

What to take from the above case is that there is a practice in economic modelling that we can call robustness analysis, which consist in deriving the same result using alternative model specifications. Economists seem to presuppose that there is value in carrying out robustness analysis, but it is unclear what this value is. In the following, we will look closer into what robustness is, introduce and dismiss the prominent proposal that robustness analysis has confirmatory value, and suggest some alternative interpretations of the value of robustness analysis.

3. Types of Robustness

Robustness is a popular concept both in science and philosophy of science. A quick scan over the literature reveals however that the term 'robustness' is used to refer to a variety of things and is said to serve a variety of functions, and it will be useful to get a sense of what the term may refer to before tackling the problem of robustness in modelling. For this purpose, this chapter develops a taxonomy of types of robustness, building up on Woodward's (2006) taxonomy of four kinds of robustness.

Robustness clearly has something to do with stability, and is sometimes used interchangeably with that term. We may for instance speak of a chair or another piece of furniture as being 'robust', or of somebody being in 'robust' condition again after spells of sickness. More specifically, what is often meant is not merely stability, but stability under change, or when the thing that remains stable is somehow interfered with. A person is in robust health if nothing makes them sick easily, and a chair is robust if it doesn't break under heavy stress.

In scientific contexts, when the term is used, we can also usually identify a thing which is said to be stable, and the interferences or changes under which that thing stays stable. For instance, take this extract from the introduction of a research paper in molecular biology:

> *"We study […] molecular homeostasis—how cells achieve a robustness to perturbations in their environment or in their internal molecular composition."* (Hartwell 2004, p.774)

What is stable here is the condition of a cell, and the changes under which this condition is stable are perturbations in cell environment and molecular composition.

Now contrast molecular homeostasis with the following use of the term 'robustness', this time from a research paper in earth science:

> *"Our results are robust and independent of the mantle model used to correct the data."* (Beghein and Trampert 2003, p.552)

Here what is judged stable is the result of a scientific investigation, and the changes under which that result is stable are changes in the model used to derive the result from a set of data. What these two uses of the term 'robustness' have in common is that there is something said to be stable, and some changes that thing is stable under; Yet, they differ in what the stable thing and the changes are exactly. These are the things we should ask about when distinguishing between different types of robustness, as I will do in the following.

3.1. Overview

Woodward (2006) provides one of the most useful and widely cited taxonomies of different notions of robustness as they are used in economics. He is motivated by the thought that authors frequently fail to distinguish between different notions of robustness where they should: Different notions of robustness differ in their normative credentials, i.e. what use they are in our scientific investigations, and in what is required for their successful deployment. I want to take Woodward's taxonomy as a starting point, and extend some of his notions of robustness. He distinguishes four kinds:

- **Causal robustness:** The stability of causal relationships under interventions
- **Measurement robustness:** The stability of a measurement result when different methods of measurement have been used
- **Inferential robustness:** The stability of an inference from data to a hypothesis when different auxiliary assumptions are used
- **Derivational robustness:** The stability of the derivation of a theoretical result under different assumptions

Even though this is his motivation, Woodward himself does not offer any instances where different notions of robustness have been mixed up to the detriment of a philosophical or scientific argument. In the next chapter we will see that this has been the case in debates about robustness analysis in theoretical modelling, where interpretations of robustness analysis as inferential robustness and as measurement robustness are often not distinguished, and it is therefore obscured what is required for a robustness analysis to be successful.

What I would like to do in the following is to introduce and elaborate Woodward's notions of robustness and their normative credentials and requirements, making a number of modifications, which will be summarised in section 3.4. Most importantly, I would like to superimpose a distinction between what I want to call *Robustness in the Target* and *Methodological Robustness* onto Woodward's. The latter encompasses measurement robustness, inferential robustness and derivational robustness, since what all of these have in common is that they concern the stability of the output from a scientific investigation, not the stability of some property or relationship in a real world target of interest, which is what robustness in the target is all about.

To illustrate this more fundamental distinction between methodological robustness and robustness in the target, consider the two quotations from above.

> *"We study […] molecular homeostasis—how cells achieve a robustness to perturbations in their environment or in their internal molecular composition." (Hartwell 2004, p.774)*

> *"Our results are robust and independent of the mantle model used to correct the data." (Beghein and Trampert 2003, p.552)*

In the first case, robustness is a property of the thing we are trying to study, something that is 'out there' in the world, namely cells that are subject to real changes in their environment. In the second case robustness is a property of the outputs of our scientific methods, and the

changes that the result is stable under are changes in the methods used or in the ways these methods are used: In this case, a result is stable when different models are used in combination with a set of data.

Most philosophers of science who have written on robustness have been more concerned with robustness relating to our scientific methods. In that case, on the most general level, we may perhaps say that a result is 'robust' when it is stable when various ways of determining that result are applied. However it is important to see that the first use of the term robustness is also a common and important one in science – and as we will see, one that helps understand some of what economists aim to do when they are concerned with methodological robustness. It is not always obvious whether scientists want to use a variety of models to confirm one hypothesis about a target, or whether they use this variety to confirm a variety of hypotheses that combine to teach us about a robustness property of the target. Chapter 7 will argue that it is often the latter that economists are trying to do when they use robustness analysis.

3.2. Robustness in the Target

Having introduced a notion of robustness in the target can help us understand these different ways in which robustness analysis can be used. When biologists talk of a cell's condition as robust, or psychologists write on robust character traits, or when economists claim that a country's economy is robust, then they are talking about a property of the system they wish to study – such as organisms, human behaviour and economies. As is customary in the literature on modelling, let us call these systems scientists are interested in learning about the 'target' system. Often these target systems are systems we think of as being 'out there', in the 'real world'. However for our purposes, it does not matter whether we think the target is 'real' or not. The important distinction is that between the thing we are trying to study and learn something about, and the scientific tools we use to do so. Here's another example of talk of robustness related to the target of investigation from an issue of *The Week in Science* (2000):

> *"The viable development of quantum computers will depend on the implementation of procedures to overcome the problem of decoherence, where the superposition of the quantum states is lost due to disturbances from the environment. Recent theoretical work has suggested that the existence of decoherence-free subspaces can be created—a particular subset of quantum states can be chosen that will be robust to certain perturbations and not decohere."* (p.405)

The target of investigation in this case is the world of quantum mechanics, and especially its application to computing. Quantum computing uses the superposition of quantum states to more effectively encode information than ordinary computers. The research referred to in this passage addresses the problem of decoherence - that superposition may be lost due to the influence of the environment. The scientific advance is that theoretical research has shown that there are subsets of quantum states that are 'robust' in the sense that they do not decohere. The claim is that scientific investigation, in this case theoretical work, has taught us about a property of the target of investigation: We have learnt something new about quantum

states, namely that some of them can be robust to perturbations. Still, the correct interpretation of quantum mechanics is very much an open question (see Ismael 2009), and it is controversial whether quantum states could be thought of as 'real' and in what sense - The important thing that makes this kind of robustness talk different from robustness related to our scientific methods is not that robustness here is a property of something 'really there', but that it is a property of our system of investigation, the target. By robustness in the target, then, I mean stability of some property of the target under changes within the target system.

A special case of robustness in the target is what Woodward, in his 2006 paper on four varieties of robustness calls 'causal robustness'. This is essentially the idea that a cause that is operating in the target system we are interested in is stable under some changes. It could be stable when there are disturbances, such as other causal factors operating on a variable of interest. Or it could be stable under different kinds of background conditions – for instance, the tendency of GDP to fluctuate with the seasons in a certain pattern may be stable under different conditions, such as whether the country is in a recession or in a boom, or whether the government is has adopted laissez-faire or interventionist economic policy, etc.

Note that not all cases of robustness in the target need to be cases of causal robustness: Robustness in the target may also concern the stability of things that are not causes, such as the stability of certain biological traits under different conditions, or, to add an economic example, the stability of a market equilibrium. We can hence understand the notion of robustness in the target as a generalisation of Woodward's causal robustness.

Guala and Salanti (2002) introduce a similar notion of causal robustness to Woodward's. Their paper, which introduces a taxonomy of three kinds of robustness, is explicitly concerned with theoretical economic models that represent causal mechanisms, of which they would count Banerjee's model as one. For them, causal robustness is a property of the true causal mechanism at work in the target, and not a property of a model or a group of models.

Causal robustness is similar to what Cartwright (1999) calls a 'capacity'. Capacities, for Cartwright, are causes which have a stable effect over at least some range of situations, and which produce their effect ´potentially´ - by which she means that, if present, they always try to produce that effect, and always do in fact contribute something to an overall outcome. Cartwright argues that thinking about capacities is a better way to do science, and a better way to do economics, than thinking about laws. Essentially, this is because she thinks that causes are prior to laws. Causal powers, more explicitly capacities, are part of the nature of things, whereas laws are not – the empirical regularities we observe are just the result of the interaction of more or less stable causes. If Cartwright is right, then finding out about causal robustness is of foremost importance in science.

Even if we do not share Cartwright's metaphysical convictions, knowing that a cause is robust in a target system of interest is useful for a number of reasons. Firstly, knowledge about stable causes can help us make extrapolations – that is, transfer knowledge about one kind of situation to another. One big question in the methodology of experimental economics is whether we can transfer experimental results to contexts outside of the laboratory, potentially contexts which have a much richer causal structure than the isolated interactions studied in

economic experiments (see for instance Guala 2005). Here we want to extrapolate from knowledge about an experimental scenario to real, often more large scale economic interactions. Or imagine we have studied the effects of a range of labour market policies in France. Can we use our knowledge about France to say anything about Germany? Here we would like to extrapolate from knowledge about one domain to another. Knowledge about causal robustness can help us here: If we are certain that we are dealing with a capacity, or a robust cause, then we can be sure that our results will carry over.

Secondly, causal robustness, or robustness in the target in general, restricts the amount of information we need to have about the state of the target system in order to make a prediction. For instance, given what the study from molecular biology quoted above shows about the robustness of cells, we do not need to know much about a cell's environment to make a prediction about its state.

Lastly, on some accounts of causation (for instance Woodward 2003), we need a form of causal robustness, namely invariance under interventions, in order to even learn about causes, or to even be able to speak of something as a cause. For Cartwright 2009, the fact that we need invariance for causal inference and the fact that stability allows us to extrapolate and to predict are connected problems: We need invariance to learn about causes that we can use – for instance to make predictions about policy. Robustness in the target, and especially causal robustness, is hence often a useful thing to know about, and something scientists have an interest in investigating. It is quite a different thing, however, from the robustness scientists have in mind when they speak about the application of scientific methods, and the results ensuing from it.

This thesis will be mostly concerned with methodological robustness, as it will be discussed in the following. Still, it is important for us to be aware of the notion of robustness in the target, even if we cannot go into the details of what robustness in the target, for instance in the form of invariance, is used for. This is because one reason why we might be interested in using alternative methods to determine a result, especially alternative models, may be that we think these alternatives capture different aspects of the target system of interest, and that we can hence learn about the robustness of a property of the target when we use these alternative means of determination. This is an interpretation of the use of methodological robustness which is overlooked in the main literature this thesis deals with, and which I will argue can help explain much of the practice of robustness analysis in economic modelling. But when philosophers employ notions of methodological robustness, they are usually interested in using a variety of means of determination to confirm one single hypothesis about a target, or to make one kind of inference. Robustness is seen as something which occurs wholly in the employment of scientific methods of investigation.

3.3. Methodological Robustness

As we said above, often the term 'robustness' is applied to the output of a tool of scientific investigation, such as an experimental, measurement or modelling result. Here are two more

examples, one from meteorology, and another concerning an economic model of disaster avoidance:

"Our results are robust with respect to uncertainties in model estimates of anthropogenic climate fingerprints and natural variability, down-scaling method, and the choice of univariate or multivariate D&A analysis." (Barnett et al. 2008, p.1080)

"The results are robust to uncertainty about the values of the disaster probability and the equity premium, and can accommodate seemingly paradoxical situations in which the equity premium may appear to be infinite." (Barro and Jin 2011, p.1567)

When robustness refers to the stability of an output of a scientific method, it may either be the case that we use the same kind of method to determine the same result several times, but apply it slightly differently – this is the case in the two examples given above, where the researchers use one model, and vary some of the assumptions made therein. Or it may be the case that we use different kinds of methods, or different sources of evidence to determine equivalent results. Roughly, this is the distinction between what Woodward (2006) calls 'inferential robustness' and 'measurement robustness', which I want to introduce in the following. Both of these kinds of robustness have in common that they have been said to be highly confirmatory of the scientific result in question, given the right conditions hold.

3.3.1. Measurement Robustness

One of the classic examples of the alleged confirmatory power of robustness in science is Jean Perrin's estimation of Avogadro's number (the number of molecules in one gram of hydrogen) using thirteen different experimental and observational methods. This variety of methods produced the same result throughout, and it seems plausible to say that this warrants high confidence in the measurement being correct. It is remarkable and surprising that all these different ways of determining the number arrived at the same result. And why would all these different methods agree, unless the result was in fact correct? The correctness of the result appears to be the best explanation of the remarkable agreement between measurements.

This is what Woodward calls 'measurement robustness': Several different ways of measuring a quantity produce the same result. What is stable here is the result of measurement, and what is varied are the methods of measurement. The power of this kind of robustness is probably the least controversial amongst the varieties of methodological robustness, although general arguments why this is the case are still lacking (see Stegenga 2009). What does seem to be necessary for our judgement that increased confidence is warranted is that the different methods of measurement are independent of each other, and hence not likely to share the same kinds of biases. The argument then is, roughly, that it would be a remarkable kind of coincidence if the different methods agreed unless the measurement was correct: the different methods would just have to have happened to be biased in the same way by pure chance. Therefore, we can have high degrees of confidence in the measurement result.

In fact, agreements between different procedures of measurement of this kind have been thought to be so remarkable that they can not only justify concluding from robustness of a

measurement result that we have measured correctly, but also concluding that some kind of realism about the thing we are measuring is justified (see Cartwright 1983).

3.3.2. Variety of Evidence – Generalising Measurement Robustness

While Woodward sticks to speaking about measurement, we can expand this kind of argument to sources of evidence in general, and not just measurement in the common sense understanding. This is the idea of robustness as the agreement of diverse and independent sources of evidence. In this case we are trying to use evidence to confirm some hypothesis, but the evidence is uncertain. When we have a variety of independent sources of evidence, then again the argument is that when a number of independent sources of evidence support the same hypothesis, our confidence in it should increase – after all this agreement would be a remarkable coincidence if the hypothesis wasn't true. What is stable in this case is the hypothesis that is being supported, and what it is stable under is the presentation of a variety of evidence.

This idea has been formalised in Bayesian epistemology, and is known as the "Variety of Evidence Thesis" (see for instance Bovens and Hartmann 2003). We can think of measurement procedures as sources of evidence supporting some hypothesis about the measured quantity, and hence look at measurement robustness as a special case of variety of evidence for a hypothesis. While Woodward does not make a connection to the literature on the variety of evidence thesis, considering the Bayesian rationale for why the agreement of independent sources of evidence provides a strong form of confirmation is illuminating. The following box provides an illustration of the Bayesian rationale for the variety of evidence thesis using a simple numerical example.

Box 1: The Bayesian Rationale for the Variety of Evidence Thesis

We assume that we have a prior degree of belief in the truth of some hypothesis of interest H that can be expressed by a probability Pr(H). Further, we have a degree of belief in how likely some piece of evidence E is, in the case that H is true, Pr(E/H) and in the case that H is false, Pr(E/notH). From this, we can calculate an overall degree of belief in observing the piece of evidence

$$Pr(E) = Pr(E/H)Pr(H) + Pr(E/notH)Pr(notH).$$

Bayes' theorem tells us what our belief in the hypothesis given the evidence should be:

$$Pr(H/E) = Pr(E/H)Pr(H) / Pr(E)$$

Upon observing evidence E, Bayesians tell us to update our beliefs using this rule: We should replace our degree of belief in H by the belief in H conditional on E – The new Pr(H)* equals Pr(H/E).

To now give a trivial example of how variety of evidence has confirmatory value within this framework, imagine we have two pieces of evidence E1 and E2 which are independent of each other in the sense that given H is true, and given H is false, knowing that one piece of evidence has been observed, our belief in observing the other is not

changed. In other words, E1 and E2 are conditionally probabilistically independent. This implies that

Pr(E1/E2,H) = Pr(E1/H), and that
Pr(E1&E2/H) = Pr(E1/H)Pr(E2/H), and similarly for notH.

In this case, Bayes Theorem applied to the evidence E1&E2 becomes:

Pr(H/E1&E2) = Pr(E1&E2/H)Pr(H) / (Pr(E1&E2/H)Pr(H) + Pr(E1&E2/notH)Pr(notH))
= Pr(E1/H)Pr(E2/H)Pr(H) / (Pr(E1/H)Pr(E2/H)Pr(H) +
Pr(E1/notH)Pr(E2/notH)Pr(notH))

Now, to look at a numerical example, take the case where

Pr(E1/H) = Pr(E2/H) = 0.8;
Pr(E1/notH) = Pr(E2/notH) = 0.1; and
Pr(H) = 0.5.

In this case, our new degree of belief in the hypothesis if only one piece of evidence is observed is:

Pr(H/E1) = Pr(H/E2) = 0.8*0.5 / (0.8*0.5+0.1*0.5) = 0.4 / 0.45 = 0.8888…

Observing both conditionally independent sources of evidence, however, gives us

Pr(H/E1&E2) = 0.8^2*0.5 / (0.8^2*0.5+0.1^2*0.5) = 0.32 / 0.325 = 0.9846…

Given these independencies and prior P(H)=0.5, we can also see that Pr(H/E1&E2) is generally greater than Pr (H/E1), for the case where Pr(E1/H) is greater than Pr(E1/notH), that is the case where we in fact think the evidence supports the hypothesis.

Of course this is only an example, and more general arguments have been offered that allow for different degrees of independence between pieces of evidence, and purport to show that the more independent the pieces of evidence are, the more support we get when we observe all the pieces of evidence as opposed to only one (see Wayne 1995). While there are some important caveats (see Bovens and Hartmann 2003), the variety of evidence thesis can be shown to hold across a wide variety of cases using Bayesian models.

It is controversial whether a Bayesian calculation is all there is to explaining why a variety of evidence, or a variety of different measurement procedures appears to be so highly confirmatory. Especially in the case of measurement robustness, which has been discussed widely outside a Bayesian context, it has been claimed that the point is precisely that we do not know how reliable each instrument of measurement is, that is, we cannot specify the likelihood of the evidence (see Woodward 2006). We just know that there may be sources or error. Finding a robust result reassures us that in fact, no error occurred, or that the individual errors of the measurement procedures did not distort our results.

Still, both the Bayesian argument, and this argument rely on the idea that it would be too much of a coincidence that independent sources of evidence agree if a hypothesis wasn't true. We may perhaps summarise these ideas of robustness, measurement robustness and the variety of evidence thesis using Wimsatt's (1981) phrase of robustness as 'triangulation via independent means of determination', since it highlights the two most important aspects of this kind of robustness: triangulation – trying to pin down a specific result by approaching it in different ways, and the required independence of the ways of determining that result. Even in the case where we cannot specify the likelihoods of the evidence precisely, because we have more fundamental doubts about the reliability of our instruments, we need some confidence that the instruments do not share the same biases in order for an argument from robustness to work. So we need to be fairly confident that the instruments are independent in the sense that they do not share similar biases. In the absence of further information, what we could go by is what has sometimes been called 'ontic independence' (Stegenga 2009), namely that the different instruments are physically different, perhaps even access the quantity measured via different causal routes. The main point for us is that, even if we are not in a position to judge whether two sources of evidence are probabilistically independent in the way we used it in the Bayesian calculation, we need some form of independence in order for an argument from robustness to be convincing. Unless we can rule out that the different sources of evidence share the same biases, the truth of the hypothesis the sources support is not unambiguously the best explanation for the agreement.

So what is central to the confirmatory power of robustness when we are dealing with variety of evidence is the independence of the different sources of evidence. This contrasts it with inferential robustness, which we will examine next, where a notion of exhaustiveness of measures tried, and found to agree, is central to the confirmatory power of robustness.

3.3.3. Inferential Robustness

Woodward identifies another kind of robustness that applies to our methods rather than the target system, namely 'inferential robustness'. Woodward argues that this kind of robustness is very different from measurement robustness, and it is in fact the target of most of his criticism. To take again the meteorology example from above:

> *"Our results are robust with respect to uncertainties in model estimates of anthropogenic climate fingerprints and natural variability, down-scaling method, and the choice of univariate or multivariate D&A analysis." (Barnett et al. 2008, p.1080)*

Here the researchers had a set of data about the hydrology of North America, but to determine any results, they needed to make assumptions about things they are uncertain about, such as anthropogenic climate fingerprints, or the right statistical model. But they claim that their results are in fact robust with regard to using alternative assumptions regarding these uncertainties.

That this kind of robustness is a good thing is a belief many econometricians hold, and this kind of robustness, when making inferences from a set of data, is what much work in econometrics

aims at. Hoover (2006), in the introduction to a symposium on robustness and fragility in econometrics in the *Journal of Economic Methodology*, puts it aptly when he writes:

> *"A pervasive idea in applied economics is the notion that empirical results are more reliable or secure if they are 'robust' (or not 'fragile'). No argument is typically given to support the notion that robust results are epistemically virtuous nor that fragile results are epistemically vicious. It just seems obvious to most applied economists. Robustness is measured against a bewildering array of variations. A researcher is happy when he or she shows that the same result can be found in different time periods, in different datasets, using different sets of variables, using different functional forms (linear or non-linear; logit or probit, etc.), using different transformations of data (levels, differences, logarithms, growth rates), using different estimation methods, using cross-sectional, time-series or panel data, and so on."* (p.159)

Let us focus on just those transformations where the same set of data is used[3]: For instance, when econometricians employ a different functional form - here economists hold that if an inference from data is stable to changes in the model used to make the inference, the inference is more reliable. This is what Woodward calls 'inferential robustness'.

To be a little more precise, assume we have one fixed set of data, D, and we try to make an inference from that data to a conclusion S, for instance about the truth of a hypothesis, or the value of a parameter. In order to do so, we need to make some additional assumptions, but we are not sure what the right or appropriate ones are. What inferential robustness, in Woodward's sense consists in is that we get the same result S under different sets of assumptions A_i. Woodward puts it as follows:

> *"Suppose […] that there are a number of different, competing possibilities A_i regarding these assumptions, and that available background knowledge provides no strong reason to prefer one of the A_i over the others. A number of writers suggest that if for each of these A_i, D supports S, this provides a strong reason for belief in S, even in the absence of information about which of these A_i is correct- S is said in this case to be robust or sturdy or insensitive to alternative assumptions A_i, given D. Call this inferential robustness."* (p.219)

Woodward, as Hoover, is specifically concerned with economics, and the literature he is referring to is what is known as 'sensitivity analysis' in econometrics, with Leamer (1983, 1985) as one of its main proponents.

The basic idea behind sensitivity analysis is that it is a good thing if a conclusion we are interested in does not depend on the precise assumptions we made when making inferences

[3] Another central concern for econometricians that concerns the stability of a result, is what has been called 'replicability'. There is an earlier philosophical literature (see for instance a mini-symposium in *History of Political Economy* (1991), and especially Collins 1991) dealing with replicability, which occurs when the same results can be produced by different economists, and/or with different sets of data. While I cannot delve into this literature fully here, my feeling is that this is best treated as a kind of measurement robustness.

from a set of data. If the conclusion is robust with regard to changing assumptions, then we gain confidence in our inference. The converse is also often asserted, namely that if a conclusion is sensitive to changing assumptions when we are not sure about what the appropriate assumptions would be, i.e. when it is 'fragile', our inference is somehow flawed, and we should be suspicious of the conclusion. This latter claim has been the focus of much criticism, including Hoover and Perez (2004), and Aldrich (2006), who argues that fragility in the sense that a result is sensitive to omitted variables is not a bad thing,. Most importantly, this fragility may just be evidence that the result we are interested in occurs only under specific circumstances in the target (i.e. it is not robust in the target in the relevant sense), not that our inference about those specific circumstances is flawed. In the following, we will hence only be concerned with the claim that (inferential) robustness is a good thing, not with the claim that fragility is bad.

Woodward stresses that inferential robustness is very different from measurement robustness, and we may add, robustness as variety of evidence. For him, the crucial difference is that no new data is added when we vary the assumptions in our inference from one given set of data. Measurement robustness and variety of evidence, in the case where we are dealing with empirical evidence, mean that different sets of data all point in the same direction. But inferential robustness occurs against the background of a given set of data. The goal is also different: When we find a result to be robust in the sense of the agreement of a variety of evidence, we gain confidence in a conclusion by combining different methods, of each of which we are uncertain, but for which we have some prior idea of its reliability. In the case of inferential robustness, we try to gain confidence an inference from one kind of instrument. Rather than using different methods to determine a result, we play around with one method to gain confidence in its reliability: We learn something about our instrument of scientific investigation, and thereby gain confidence in an inference we make with it.

When we look closer at exactly why people have thought inferential robustness to be confirmatory, another potential difference becomes clear, namely that what is crucial in inferential robustness is the exhaustiveness of the assumptions tried, whereas arguments from variety of evidence rely heavily on independence of different sources of evidence. To see this, consider how Orzack and Sober understand arguments from robustness as they are made in Levins (1966). We have a set of models M_i making alternative assumptions:

> One of $M_1 \ldots M_n$ is true.
>
> All M_i yield result R.
>
> ---
>
> Therefore, R is true.

This argument is obviously valid, but it hinges on our being certain that one of the models is 'true' and will hence give us a true result R. The requirement that we know for sure that one of the models is reliable has been held to be too strong in most contexts – perhaps a high degree of confidence is enough. Further, the suggestion that models can be true as such has been under much attack – we would have to be able to identify a model with a complex hypothesis

about the target, which is often not possible, and in any case this hypothesis would rarely be true seeing that we can never fully eliminate all idealisations from a model (see Parker 2011 for both points). Instead, Parker suggests a weaker form of this argument, which seems to be underlying what authors like Leamer have in mind when they conduct sensitivity analysis:

> It is likely that at least one of M_1….M_n indicates correctly with regard to hypothesis H.
>
> Each of M_1….M_n indicates the truth of H.
>
> ---
>
> Therefore, H is likely to be true.

What gives this argument its strength is that we are fairly confident that the set of models we looked at contains one that adequately indicates H. For the argument to work, we do not need to quantify our confidence in each model separately, or need to know whether the models share certain biases. What counts is how extensive the list of models we tried is and how confident we are that it contains an adequate one. The basic intuition is not that it would be a an incredible coincidence if the different models agreed unless the hypothesis was true, but that it would be an incredible coincidence if none of the models indicated correctly. And then, given they all happen to say the same, we can be fairly certain that they indicate correctly. The second argument relies crucially on exhaustiveness, on the belief that one of the sources must tell the truth, and does not rely on any notion of independence.

When it comes to the confirmatory value of robustness, the crucial difference between robustness as the agreement of a variety of evidence and inferential robustness turns out to be this: the confirmatory value of the first relies on independence, and the confirmatory value of the second relies on exhaustiveness.

3.3.4. Derivational Robustness

Woodward claims that there is a third kind of robustness that is methodological in nature, next to causal robustness, measurement robustness and inferential robustness, which he refers to as ‘derivational robustness’. This kind of robustness concerns the derivation of an observational result, as it often occurs in modelling: There is a phenomenon of interest that has been observed, and we use models to try and make sense of this phenomenon. In order for a model to allow us to do so, at least in economics it is customary that it needs to provide a mathematical derivation of the observational result using a set of assumptions. Derivational robustness, according to Woodward, now consists in the derivation of the same result using different sets of assumptions. This is exactly the practice we identified in the last chapter in Banerjee’s model.

Woodward claims that derivational robustness is very different from inferential robustness: For one thing, it can concern the variation of what are judged to be ‘important parameters’, while the theoretical framework stays largely the same. In inferential robustness, the idea was that only the so-called ‘background’ assumptions are varied:

> *"The two notions seem associated with very different questions and are used for different purposes. In the former case, the goal is to infer to the truth of some conclusion or the value of some parameter (not independently known) given data D and additional assumptions A_i. The claim is that inferential robustness functions as a measure of the inductive warrant D provides for the conclusion. In the latter case, in which the concern is with derivational robustness, an assumption is adopted about the value of the parameter and this is used, in conjunction with other theoretical assumptions, to derive some range of observed phenomena. Investigations are then made whether, given other values of the parameter, but the same theoretical assumptions, the same conclusions can be derived." (p.233)*

Woodward has nothing further to say on why we would want to derive the same result with models making different assumptions, and here describes the mere activity as the goal. This is unsatisfactory, given how puzzling this practice is, as we saw above in the case of Banerjee's model. Essentially, what Woodward leaves us with is that there is a further kind of robustness, different from inferential, measurement and causal robustness, which applies only to the case of theoretical modelling, with no hint to what its uses are. Robustness in modelling is just 'something else'.

In fact, I want to argue in the following chapter that there is an influential literature, reaching from Levins (1966) over Wimsatt (1981) to Weisberg (2006a, 2006b) and Kuorikoski et al. (2010) which views robustness analysis in theoretical modelling just like robustness as it occurs with other methods of scientific enquiry – namely either as the agreement of a variety of evidence or as inferential robustness. For these authors, robustness in theoretical modelling is not 'something else', or so I will argue.

3.4. Summary

This chapter provided an overview of notions of robustness in philosophy and science. To this purpose, it elaborated and extended Woodward's taxonomy of four kinds of robustness. The main advances on Woodward's taxonomy that we have made are the following:

- A distinction between robustness in the target and measurement robustness has been superimposed on Woodward's four types of robustness.
- Woodward's causal robustness has been generalised to robustness in the target, since we may be interested in the stability of properties of our target system which are not causal.
- Woodward's measurement robustness has been generalised to agreement of a variety of evidence, to take into account any kind of evidential practice, and make a fruitful connection to the literature in Bayesian epistemology dealing with variety of evidence that Woodward does not make.
- The normative credentials of inferential robustness and agreement of a variety of evidence have been elaborated: for the former, independence is crucial, for the latter, exhaustiveness of alternative assumptions tried.

- It has been noted that the notion of derivational robustness is in need of development.

The following table provides an overview of the modifications we have made to Woodward's taxonomy. The last two observations just made are captured by the two columns on the right. While we have seen that inferential robustness and robustness as the agreement of a variety of evidence aim at confirmation, and get their force from independence and exhaustiveness respectively, both the purpose and the success criteria for derivational robustness are unclear.

Type	Sub-Type	Purpose	Success relies on…
Robustness in the Target	*(Includes Woodward's Causal Robustness)*	May help with extrapolation and prediction	Existence of stable causes, phenomena, or properties of the system
Methodological Robustness	Agreement of a Variety of Evidence *(includes Woodward's Measurement Robustness)*	Confirmation – triangulation of a result	Independence of the pieces of evidence
	Inferential Robustness *(as in Woodward)*	Confirmation – making inferences more reliable	Exhaustiveness of the assumptions tried
	Derivational Robustness *(as in Woodward)*	?	?

Interestingly, the observation that the purpose of derivational robustness is left unclear by Woodward raises the same question that we encountered in the last chapter: there is this practice of deriving the same result using alternative models in economics, but we do not know exactly what the use of it is. In part II we will see that a number of philosophers of science have thought of derivational robustness either as a species of inferential robustness or of robustness as the agreement of a variety of evidence. But because part II also shows that these accounts of robustness analysis in modelling fail I will argue in part III that derivational robustness indeed aims at something else – thereby filling in for what Woodward left undeveloped.

The distinctions I made in this chapter appear to me to be the most intuitively plausible ones, and align well with the different kinds of normative arguments that are made when authors appeal to robustness. However, the real test of any taxonomy is whether it is of use in dissolving scientific problems or improving scientific practice. To demonstrate that this taxonomy can do so, the next chapter, opening part II, will illustrate how the notions of inferential robustness and robustness as agreement of a variety of evidence can help us understand debates about the confirmatory value of robustness in economic modelling, which form the main theme of part II.

PART II: Disputing the Confirmatory Value of Robustness Analysis

Part I introduced both the practice of robustness analysis in economic modelling, as well as the general literature on robustness in science. This part will bring these two themes together in that it concerns robustness as it is discussed philosophically in the literature on *modelling*. To use the taxonomy developed in the last chapter, we will be dealing with derivational robustness. As we saw in the last chapter, Woodward left this type of robustness strangely undeveloped and was not explicit on its purpose in scientific enquiry. What he was explicit about, however, was that derivational robustness is something altogether different from both inferential robustness and measurement robustness.

Chapter 4 will introduce the literature that has been concerned explicitly with derivational robustness. This literature has seen the purpose of derivational robustness to be the same as that of inferential robustness and variety of evidence arguments: namely confirmation. We will see that a variety of authors, starting with Levins (1966), have argued that robustness analysis in modelling in general can allow for more reliable inferences from models. Recently, Kuorikoski et al. (2010) have taken up this literature and argued for the confirmatory value of robustness analysis specifically in economics, and they will be our main target. By arguing that robustness analysis allows for more reliable inferences from models, their arguments connect with the problem of how to learn from the idealised models of economic theory that was introduced in chapter 1.

What chapter 4 will establish is that we can use the taxonomy developed in chapter 3 to discern two distinct arguments for the confirmatory value of robustness analysis in the literature: one that sees the models compared in robustness analysis as providing a variety of evidence for a hypothesis, and one which sees robustness in modelling as a species of inferential robustness. On neither view is derivational robustness something altogether different, as Woodward claimed. Further, Kuorikoski et al. appear to be confused between these two arguments. However, the two arguments have different requirements for their success, so it is worthwhile to distinguish them in order to assess their adequacy.

Chapters 5 and 6 will assess the two arguments distinguished in chapter 4 respectively, and show that they are unsuccessful – both at establishing that current economic practice provides confirmation through robustness analysis, and at providing a perspective for it to achieve confirmation. Ultimately, Kuorikoski et al. are mistaken in borrowing ideas on robustness applied outside of modelling – as part III will show, economists really do aim at something different when they conduct robustness analysis.

4. Accounts of Robustness in Modelling – The Alleged Confirmatory Value of Robustness

This chapter introduces the most influential literature on the benefits of robustness analysis in modelling, reaching from Levins (1966) over Wimsatt (1981) to Weisberg (2006a, 2006b). This literature has argued that robustness analysis in modelling can be confirmatory. Whereas the cited authors were concerned mostly with models in biology, more recently, Kuorikoski et al. (2010) have made this argument specifically for the case of economic models. It is their arguments that the following two chapters will show to be misguided.

In order to criticise any idea, it is good to have a clear understanding of what the view is that one is trying to argue against, and to formulate it in the most convincing and coherent way possible. The literature that I want to criticise in this part of the thesis leaves much implicit and open to interpretation. Before turning to my counterargument, I hence want to develop what I take to be the most plausible rendering of the arguments made in the literature.

What I would like to show in this chapter is that there are two distinct arguments for the confirmatory value of robustness analysis made in this literature, and that the taxonomy introduced in the last chapter can help us to understand them: The views of these authors are best understood as either aiming at a kind of inferential robustness, or at demonstrating the agreement of a variety of evidence. These authors thereby view models like sources of evidence, and think of robustness analysis as confirmatory in the same way as robustness when it comes to the agreement of instruments of measurement and the inference from data. One of the most important purposes of theoretical models in economics is indeed that we would like to use them to learn about a target, that is to make inferences from the model – just like we use other scientific tools to learn about a target. These authors think that robustness analysis can help us make better inferences from models. Contrary to Woodward (2006), for them, robustness in theoretical modelling is not 'something else', or so I will argue.

Conceiving of this literature in this way is helpful, and demonstrates the use of the distinctions we drew in the last chapter. It helps us clarify what the requirements are for robustness analysis to be successful, and draw parallels to robustness in other forms of scientific enquiry. Further, it helps us clear up some of the debate: As we have seen, arguments involving inferential robustness and robustness as the agreement of a variety of evidence take their force from different sources, which is why it is important to distinguish them. In the literature on robustness in modelling, the two kinds are sometimes mixed up, as I will show they are in Kuorikoski et al. (2010). This makes it unclear why robustness analysis is seen to be confirmatory by these authors. Another misunderstanding from not distinguishing these two accounts, which will become evident in chapters 5 and 6, results from the fact that most critics have targeted the argument that is in fact less prevalent in the literature, while the more prominent account - robustness analysis as inferential robustness - remains 'undercriticised'.

Note that while we have developed a detailed taxonomy of robustness, we have not given models and modelling the same kind of conceptual attention. My primary focus is on applying

ideas about robustness to modelling, and trying to make sense of a specific debate within the philosophical literature on modelling. The literature I deal with does not explicitly commit to a particular view on modelling, and rather than viewing it through the lens of philosophical accounts of models, I want to scrutinise it in its own right. Many interesting things will be said about modelling in the course of this exercise, and more general conclusions about modelling will be warranted in the end.

4.1. Levins, Wimsatt, Weisberg and Kuorikoski et al. on Robustness in Modelling

Let me give an overview of the most influential literature on robustness in modelling in this section, before interpreting it in terms of inferential robustness and robustness as agreement of a variety of evidence.

The biologist Richard Levins was one of the first to write about the concept of robustness in theoretical modelling, especially in his 1966 "The strategy of model building in population biology", and later in his 1968 *Evolution in Changing Environments: Some Theoretical Explorations*. He observed that theoretical biologists often used alternative models to derive the same result. What he thought they aimed to discover is "whether a result depends on the essentials of the model or on the details of the simplifying assumptions" (1966, p.423). according to him, this is an important activity because we often cannot tell whether certain simplifications are harmful or not. If models can confirm that a result does not depend on the details of the simplifying assumptions, then it provides a form of confirmation. Levins writes:

> *"[W]e 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. Hence our truth is the intersection of independent lies."* (1966, p.423)

The idea of a robust theorem has been quite influential since Levins, and is also taken up by Wimsatt (1987) and Weisberg (2006a, 2006b).

William Wimsatt is another prominent figure who has written on robustness in the context of modelling. His general idea of robustness is that "[t]hings are robust if they are accessible (detectable, measurable, derivable, definable, producible, or the like) in a variety of independent ways" (2007, p. 196). Elsewhere, as we already noted above, he characterises robustness as the triangulation of a result via independent means of determination (Wimsatt 1981), thereby picking up a term first introduced by Webb et al. (1966) in the context of social science research. 'Triangulation' is still the term most frequently used in social science when a multiplicity of methods is used to determine a result. When speaking about triangulation, Wimsatt appeals to the unlikelihood of agreement when the result to be determined isn't true – this is why triangulation is a strong method of confirmation.

Wimsatt writes about robustness in modelling more specifically as well. In Wimsatt (1987) he claims the following:

> *"A family of models of the same phenomenon, each of which makes various false assumptions, has several distinctive uses: (a) One may look for results which are true in all of the models, and therefore presumably independent of different specific assumptions which vary across models. These invariant results (Levins' (1966) "robust theorems") are thus more likely trustworthy or "true". (b) One may similarly determine assumptions that are irrelevant to a given conclusion. (c) Where a result is true in some models and false in others, one may determine which assumptions or conditions a given result depends upon." (p.32)*

Here he echoes much of what Levins has said about robustness: It is about determining whether a result depends on some fairly arbitrary modelling assumptions or not.

A more recent philosopher of biology who has argued for the value of robustness analysis in modelling and who builds up on this literature is Michael Weisberg (2006a, 2006b). He, too, thinks that robustness analysis is about the discovery of robust theorems, a notion which he cashes out more precisely in Weisberg (2006b). Here, he interprets a robust theorem as a conditional statement linking some model outcome of interest that a number of models agree on with those assumptions that have turned out to be relevant (the common structure shared by all the models). It is like a model with all the unimportant assumptions discarded. Such a conditional statement is not strictly speaking a theorem, but I will nevertheless stick to this terminology below. So for instance, if in our model, we have N assumptions A1 – AN, and derive result C, but then find out through robustness analysis that assumptions A1-A5 do not matter to the derivation of C, our robust theorem will state that:

> If A6-AN, then C

Weisberg goes on to argue that robust theorems have a "degree of confirmation", even in the absence of additional empirical evidence.

Recently, Kuorikoski et al. (2010) have taken up this literature and argued for the confirmatory value of robustness analysis specifically in economics, and they will be our main target. For them robustness analysis is what is most characteristic about theoretical progress in economics in general, and an extremely important part of economic modelling practice: Economists spend much time and effort deriving old results with ever new models. And they argue that this practice has confirmatory value.

Kuorikoski et al. begin by appealing to Wimsatt's notion of robustness as triangulation via independent means of determination, claiming they want to extend it to the case of economic modelling.

> *"Fairly varied processes or activities such as measurement, observation, experimentation, and mathematical derivation count as forms of determination. Triangulation may involve more than one of these forms (e.g. when the same result is obtained by experimentation, derivation, and measurement) or concern only one of them: the same result can be obtained by different experiments or, as in our case, by different theoretical models."* (p.542)

Then they characterise the use of robustness analysis as follows:

> *"First, it guards against error by showing that the conclusions do not depend on particular falsehoods. Secondly, it confirms claims about the relative importance of various components of the model by identifying which ones are really crucial to the conclusions."* (p. 543)

This second statement, similarly to what Wimsatt says about modelling, echoes very much what Levins has said about robustness analysis above. So for Kuorikoski et al. robustness is a form of triangulation via independent means of determination, and it can also tell us which idealisations are truly harmful, in the way Levins and Weisberg envisage.

Much is left unclear in the literature I described here, and it will be illuminating to use some of the ideas developed in the last chapter to try and develop what these authors might be aiming at. What I want to argue below is that there are two different ways of conceiving of robustness analysis hidden in this story, and in Kuorikoski et al.'s paper in particular: Namely conceiving of robustness in modelling as a form of agreement of a variety of evidence, and conceiving of it as a form of inferential robustness. I will delineate the two approaches, and also show why they should be distinguished.

But before we delve into distinguishing the two views on robustness analysis, note that by considering robustness analysis to be confirmatory, the authors we discussed regard models to be sources of evidence - they can help us confirm claims about a target. Nevertheless, they are imperfect sources of evidence, and inferences from models can be improved via robustness analysis. The literature I discussed is not explicit about any of this - but seeing that it is not entirely obvious how a model like Banerjee's could be thought of as a source of evidence, I want to do two things in the next section, before turning back to making sense of the literature on robustness in modelling. Firstly, I will introduce a schema of what it might mean to make an inference from a model. Secondly, I will indicate what kind of problems we may face when trying to make inferences from models, problems that robustness analysis has been thought of as helping us to overcome. This will help us making sense of the literature just described.

4.2. Making Inferences from Models, and the Problem of Unrealistic Assumptions

So by viewing robustness analysis to be confirmatory, the authors I mentioned must view models to be something like sources of evidence. Kuorikoski et al. explicitly write:

> *"Modelling can be considered as an act of inference from a set of substantial assumptions to a conclusion."* (p.561)

Note that this means that models are not viewed merely as interpretations of theory as they are, for instance, in the syntactic view of theories (see for instance Hempel 1965). They are instruments of scientific investigation in their own right. At the same time, they are not empirical models either, as models in econometrics are – they are not made primarily to facilitate inferences from a set of data. Perhaps the view of models that best fits this approach is Morgan and Morrison's (1999) view of models as autonomous agents, where models are *"partially independent of both theories and the world"* (p.10). There are also accounts of modelling that see models more specifically as inferential devices, or as instruments of measurement (see for instance Boumans 2005).

The authors I referred to proceed without taking a more specific stance on what models are, and we will see that much of the criticism we rehearse in the next few chapters can proceed in such a way. Let me note, however, that what will come out of the investigation of this part of the thesis is that the authors whose arguments I scrutinise have implicitly taken a view on what models are, and how models are made up of assumptions that is highly implausible. An implication of this is that to understand the practice of robustness analysis, we would have to have a better understanding of what models are and how they function. But for now, the most important thing we do have to be more specific about are the kinds of inferences that can be made from models.

Sugden (2000) provides a useful overview of the kinds of inferences we might want to make from models. He distinguishes three kinds of inferences made from models. The first kind is concerned with explanation:

> E1. In the model, R is caused by F.
>
> E2. F operates in the target.
> E3. R occurs in the target.
>
> *Therefore, there is some reason to believe*:
> E4. In the target, R is caused by F.

The second kind of inference concerns prediction:

> P1. In the model, R is caused by F.
>
> P2. F operates in the target.
>
> *Therefore, there is some reason to believe*:
>
> P3. R occurs in the target.

The third kind is what Sugden calls "abductive":

> A1. In the model, R is caused by F.
>
> A2. R occurs in the target.
>
> *Therefore, there is reason to believe*:
> A3. F operates in the target.

Note that this scheme requires a causal interpretation of models: We have to identify a cause and an effect in the model, and the very same cause and effect are then thought of as materialising in the target. This claim is somewhat controversial but I will stick to it here, especially because the literature on robustness, as we already saw in the case of Guala and Salanti (2002), largely employs a notion of models as describing causal mechanisms.

As noted in the introduction, the debate about what might license these kinds of inferences is one of the most longstanding in the philosophy of economics. It is still very much an open question whether, what and how we can learn from economic models of the type described above – that is, how models can tell us anything about 'the real world' when they are so different from this world in many respects. Following Friedman (1953) this debate has been centred on the problem of 'unrealistic' assumptions[4]. The problem is that mathematical models like Banerjee's are very simple compared to the real world phenomena they purport to apply to (the target) – they lack much of the detail of the real world phenomena. Furthermore, many of the assumptions made in the models are literally false of the target. For instance, imagine we want to apply Banerjee's model to herd behaviour in financial markets. We may miss a lot of detail: About markets, about varieties of financial products, and about the decision processes of the banks and firms that participate in these market. Many of the assumptions of the model are false of the target: There is never a continuum of options in a real life market - Options are finite. Agents are usually not fully rational, and common knowledge of the structure of the interaction, in the strict sense, is normally unattainable. Agents may not all have the same utility function. Even if they all care about money only, some may be risk-averse. Decisions in financial markets are frequently made without knowing about all of the actions previously undertaken by rival traders. And the distribution of information is probably much more complex than described in the model: The signals received by the individual agents may not be completely private, or they may be correlated rather than independent; they may be vague.

But given all this, how can we say that models like Banerjee's could tell us anything at all about real world financial markets? Unrealistic assumptions seem to stand in the way of making inferences from models that we can trust. There are several standard accounts of economic modelling that attempt to give an answer to this question. Suffice to say, while the fundamental debates about unrealistic assumptions are unresolved, most philosophers of economics do in fact believe that we are sometimes licensed to make the kinds of inferences that Sugden identified above. Still, unrealistic assumptions remain a worry. Hence, if robustness analysis could in fact, like the authors cited above believe it can, help us with making inferences from models more reliable, this would be a great advance in the on-going debate on how we can learn from simple economic models.

We are now ready to make sense of the literature described in 4.1.: We have an idea of what it might mean to make inferences from economic models and what the problems in doing so

[4] It may be slightly confusing to speak of the 'realism' of an assumption when we have said that we need not be committed to the reality of the target we are trying to learn about. Still, this term is very common in this literature, so I will employ it. All I have in mind here is whether the assumptions made in a model are true of the target system or not, whether they describe it accurately.

may be, and we are equipped with a taxonomy of what robustness has been said to achieve in the broader scientific context, developed in chapter 3. I claim that there are two distinct arguments that can be made for why robustness analysis can make inferences from models more reliable and hence confirm claims about a target. Both can be found in the literature, but they are sometimes not clearly discerned.

4.3. Robustness in Modelling as Agreement of a Variety of Evidence

Some of the proponents of the epistemic use of robust modelling results in theoretical economics talk as if models were just like any other source of evidence that can combine to deliver a robust result supported by a variety of evidence. Wimsatt's notion of triangulation seems to be just another way of expressing the idea that we gain confidence when independent ways of determining a result, such as different measurement procedures, agree because the correctness of the result is the best explanation for the agreement – i.e. the variety of evidence argument that relies on independence of sources of evidence that we in chapter 3. Wimsatt himself is not concerned specifically with modelling when he speaks of triangulation via independent means of determination. But when Kuorikoski et al. claim that they want to develop this notion of Wimsatt's they apply precisely this idea to the case of modelling:

> *"Independent ways of determining the same result reduce the probability of error due to mistakes and biases in those different ways of arriving at the result. Wimsatt generalises this principle to all forms of fallible ampliative inference." (p. 544)*
>
> *"In the following sections we generalise this principle from experiments and measurements to theoretical modelling. In effect, we treat theoretical models as forms of determination."* (p.545)

If we conceive of models in this way, we think of them as sources of evidence with some probability of error, just like a measurement procedure – where the sources of error are potentially introduced by the presence of unrealistic assumptions. If we can rule out that the different models share the same biases, then we can make an argument from variety of evidence like the ones identified in the last chapter: It would be a surprising coincidence if the models all agreed and the result they all pointed to wasn't correct.

What the models compared in a robustness analysis are supposed to be evidence for is sometimes not so explicitly stated. Using Sugden's scheme, we could say firstly, that models serve as evidence that a certain result from the model occurs in the target. This is what Sugden calls prediction. In Banerjee's case, maybe we could say that his model, and altered versions of it are evidence that herd behaviour, the phenomenon that everybody ends up choosing the same thing despite some having signals to the contrary, will occur in the target of interest. Secondly, we could say that given the result of a model occurs in the target, the model is evidence that the factors that cause the result in the model are present in the target. This is what Sugden calls abduction. So in Banerjee's case, we would take a situation where we have

observed that everybody chooses the same in our target, such as everybody choosing the same restaurant. Then a model which tells us that a certain kind of informational structure causes this result would act as evidence that this informational structure exists in the target. Thirdly, we could think of a situation where we know both that a particular informational structure as well as the phenomenon of herd behaviour occur in the target, and of the model, in which the herding result is caused by the informational structure, as evidence that the informational structure causes herd behaviour in the target. This is what Sugden calls explanation.

According to this view, then, robustness analysis could provide us with a variety of evidence for these different claims. In each case, the requirements of what needs to be stable across the models are slightly different: In the first, only the result (herd behaviour) needs to be shared by all the models. In the second and third, both some important causing factors as well as the result need to be shared. In all the cases, however, the models need to in fact all be evidence for the respective claims. As we have said, there is much disagreement how models can serve an evidential role. Still, it seems to be generally agreed that models need to be similar to their targets in relevant respects, or at least that such similarity is generally a good thing. In that case the models compared in a robustness analysis probably all need to share certain assumptions that make them relevantly similar to the target, beyond the results we require them to agree on.

In chapter 3 we said that the most important requirement for a variety of evidence argument to be successful is that the different sources of evidence are independent in the sense that they do not share the same biases. Hence, this is the standard by which this argument for the confirmatory power of robustness analysis should be judged.

The challenge to this view is then to show that the different models compared in a robustness analysis are in fact independent sources of evidence, in the sense that they do not share the same kinds of biases. If we want to apply the Bayesian rationale, we have to think of independence as a model's delivering the robust result given the result points to a true hypothesis as probabilistically independent of another model delivering that result. Given the fundamental worries about unrealistic assumptions just described, it may be hard to say anything about the likelihood of a model producing a certain result, given the hypothesis it points to is true, and hence to get a Bayesian argument off the ground. But maybe we could have some vaguer idea of whether models share similar biases.

Problems with this conception of robustness analysis in modelling will be explored in the next chapter. Interestingly, this argument has been the main target of the critics of the idea of getting confirmation out of robustness analysis. However, most of the authors discussed in 4.1. appear to defend a different argument, one that sees robustness analysis more like seeking inferential robustness. This argument will be discussed in the following.

4.4. Robust Theorems and Inferential Robustness in Modelling: How Robustness May Help with Unrealistic Assumptions

What I want to argue here is that Levins, Weisberg, and Wimsatt when he is speaking specifically about modelling, have a different approach to robustness in mind, which is much more like inferential robustness as we characterised it in the last chapter. Kuorikoski et al., too, follow this approach in much of their paper, without distinguishing it from the argument described in the last section.

To recapitulate, inferential robustness concerns cases where we want to make an inference from a set of data, but have to make some auxiliary assumptions to do so. Showing that a result is robust to changes in auxiliary assumptions is supposed to give us confidence in our inference from a set of data by showing that it did not depend on the precise auxiliary assumptions made. What we have argued, in chapter 3, gives appeals to inferential robustness its power is the exhaustiveness of the assumptions tried: our being relatively confident that one set of auxiliary assumptions that allows for a correct inference is amongst the options tried. The power of inferential robustness does not rely on a notion of independence.

Now take the following quotation from Levins again:

> *"[W]e 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. Hence our truth is the intersection of independent lies."* (1966, p.423)

Not just in the writings of Levins, but also in Wimsatt and Weisberg, robustness analysis can lead to the discovery of 'robust theorems', which is the central notion in their idea of robustness: it is a conditional statement linking all the parts of a model that have been shown to matter, that part which is common to all the models, with a common modelling result. This robust theorem is then thought of as being more reliable. Unfortunately, however, it is left rather vague what exactly the *epistemic* benefit is to finding out that a model 'core' is driving a result: After all, we are just finding out about a property of our model. The best argument, I claim, involves interpreting robustness analysis as a kind of inferential robustness.

Note first that what Levins is saying sounds very much like inferential robustness in that he is interested in finding out that certain assumptions are irrelevant to the derivation of a result. Consider Levins (1966) writing that robustness analysis helps us discover "whether a result depends on the essentials of the model or on the details of the simplifying assumptions" (p.423). Only in this case, we are not concerned with making inferences from a set of data. Still, the logic these authors appeal to seems to be very similar: Some assumptions are 'essential' - the common 'biological assumption' in Levins case, that then forms the core of the robust theorem. But there are other assumptions that we are worried about, and finding out about their irrelevance gives us more confidence in an inference. The 'essential' assumptions

seem to be playing a role similar to the data, and the other assumptions, the 'details' of the model, play a role similar to auxiliary assumptions.

To make this connection to inferential robustness clearer, it is informative to view the arguments Levins and others make in light of the debate on unrealistic assumptions. As we have just seen, there are usually a number of assumptions we are worried about because they diverge from what is true of the target. For the authors just mentioned, the idea behind showing the robustness of some result is that it can help us lose some of our worries about particular unrealistic assumptions. By showing the robustness of a modelling result with respect to changing an unrealistic assumption, we show that the result does not depend on this assumption – the assumption is unimportant for the modelling result we are in fact interested in. This should increase our confidence in our ability to learn from the model. What is implicit in this is that there are certain assumptions we are not worried about, namely the core assumptions. What is not said, but helps to draw a connection to inferential robustness, is that we are not worried about them because we take them to be instantiated in the target.

A distinction is sometimes made between tractability assumptions, Galilean idealisations, and substantive assumptions in modelling. Kuorikoski et al. (2010) rely heavily on this classification, and Guala and Salanti (2002) base their taxonomy of model robustness partly on a similar classification. Along with this classification comes a vision of models as describing a causal mechanism that is supposed to explain a real world phenomenon of interest. Substantive assumptions are those assumptions which are supposed to represent the causal mechanism in the target system, or else those features of the target system that are meant to explain some phenomenon of interest. Galilean assumptions are those assumptions which isolate the causal mechanism from interfering forces: They assume away disturbances. Tractability assumptions are assumptions necessary to derive a result mathematically, or simplifications and approximations that make a model more tractable.

When we view models in this way, we want at least the substantive assumptions to be true in the target. Now what Kuorikoski et al. appear to argue is that robustness analysis tells us that the true assumptions are relatively more important than the false ones in making an inference, which makes the model more 'truthful' as a whole, and hence presumably allows for more reliable inferences. The best way to flesh this out is to say that when we make an inference with a model, we always know some part of the model to be true. And in that case, our knowledge of some assumptions being true can be seen as analogous to having data in the case of inferential robustness. To see this, let us consider the kinds of inferences Sugden talks about.

Take for instance Sugden's second type of inference, concerning prediction.

P1. In the model, R is caused by F.

P2. F operates in the target.

Therefore, there is some reason to believe:

P3. R occurs in the target.

Here we have a model with substantive assumptions including F, as well as a number of tractability and Galilean assumptions, which we use to derive R. Now we know that F is 'true' in the target (P2), and we want to make the inference that in that case R is also true in the target. We are worried, however, about all the other assumptions in our model. Robustness, so the argument goes, can make us lose some of these worries and gain confidence in our inference from our knowledge that F is the case in the target to the prediction that then R is also the case. Our knowledge that F is the case can be thought of as playing the role of the data, whereas the other assumptions in the model are the assumptions needed to make the inference, but some of which we are uncertain about. But this is exactly what inferential robustness, discussed in the last chapter, consists in: Losing worries about auxiliary assumptions needed to make inferences from data.

The argument works similarly for the other kinds of inference: In the case of explanation, i.e. the case where we want to confirm that the factors causing the result in the model represent the factors that cause the phenomenon of interest in the target, we know that the substantive assumptions as well as the result of the model are true in the target. From this 'data', we want to use the model to make the inference that the substantive assumptions describe what caused the phenomenon of interest, but we are worried about the other assumptions of our model. In the case of abduction, our only data is that the result occurred in the target, and we want to conclude from that that the substantive assumptions in the model are true.

What all of this aimed to show is that the best argument for why finding out about robust theorems in the way that Levins, Wimsatt, Weisberg, and in some instances Kuorikoski et al. envisage them can offer a degree of confirmation involves seeing robustness analysis as aiming at inferential robustness as we characterised it in the last chapter. But in that case, it is not independence, but the exhaustiveness of alternative assumptions tried that makes appeals to robustness successful: We think that an inference is reliable because we are fairly certain that one of the models tried out must have indicated correctly. Only if we have tried out an exhaustive list of assumptions can we say that a particular assumption is irrelevant to a result and will not have introduced an error. Showing that this requirement is fulfilled is the main challenge to this view of robustness analysis.

Along the lines of inferential robustness, as a general argument for the epistemic use of robustness in theoretical modelling, we can formulate the following: We want to use model M to make an inference about a target system. We are particularly interested in one result of the model, C. Depending on what kind of inference we want to make, this may be a modelling result, like the fact that herding occurs, or the fact that some substantive assumptions cause that result. We are worried about our ability to use the model for its purpose because of some unrealistic assumptions U1…N.

1) If we find out that one of the assumptions we are worried about Ui is unimportant for the derivation of C, then we gain confidence in our inferences from M.
2) Robustness can teach us that an assumption Ui is unimportant for the derivation of R.
3) Hence robustness can increase our confidence in our inferences from M.

This argument does not explicitly mention robust theorems, but the idea is intimately linked to it. A robust theorem is just the model, with an unimportant assumption dropped. So we could formulate premise 1 as saying that we have more confidence in an inference from a robust theorem with Ui dropped, than we have in an inference from the original model. The original model and the robust theorem share all the elements that are important for our inference, namely the substantive assumption and the model result of interest.

This argument will be the target of chapter 6, where both premises will be shown to be problematic – most importantly for failures of exhaustiveness.

4.5. Two Conceptions of the Confirmatory Value of Robustness – Why They Should be Distinguished

What we have seen is that some of what is said in the literature on robustness in theoretical modelling matches very neatly onto robustness as agreement of a variety of evidence, and some of it expresses the core ideas of inferential robustness, as we characterised these notions in the last chapter. In particular, Levins' and Weisberg's accounts, Wimsatt when he is talking specifically about modelling, and much of Kuorikoski et al.'s paper are best understood as aiming at inferential robustness. Kuorikoski et al. appealing to Wimsatt's general notion of robustness on the other hand, have an idea of robustness as agreement of a variety of evidence in mind. But why is it important to distinguish between these?

As Woodward also stresses, the reason lies in the fact that these approaches have different normative credentials. While both of these approaches argue for the confirmatory value of robustness analysis, they do so in different ways: As highlighted in the last chapter, for one the crucial notion is independence, and for the other it is exhaustiveness. The first approach requires us to look at models as independent pieces of evidence for a hypothesis. When they all point in the same direction, provided they are indeed independent and do not have shared biases, the best explanation for their agreeing is that the hypothesis is true – which justifies increased confidence that the hypothesis is indeed true. The second approach proceeds via the identification of some assumptions as relevant and others as irrelevant for a certain modelling result. When assumptions we are not sure about are irrelevant to a modelling result, while the ones we know to be true are, then this is said to license increased confidence in the model. What is crucial here is that we are indeed confident that we have shown some assumptions to be irrelevant, which requires that we have tried out a sufficiently exhaustive number of alternative assumptions – that we are sufficiently confident that we have tried out one that is adequate.

So if we want to show that there is confirmatory value in robustness analysis, we could try both approaches to robustness – but depending on what we choose, we will need to demonstrate different things: that models do not share biases in the one case, and that we have tried enough alternatives to allow for concluding that an assumption is irrelevant in the other. We need not necessarily show both.

Kuorikoski et al. do not make this distinction, as already becomes apparent in their abstract, in which they appeal both to triangulation, and to the idea that robustness analysis tells us which idealisations are truly harmful:

> *"We claim that the process of theoretical model refinement in economics is best characterised as robustness analysis: the systematic examination of the robustness of modelling results with respect to particular modelling assumptions. We argue that this practise [sic.] has epistemic value by extending William Wimsatt's account of robustness analysis as triangulation via independent means of determination. For economists robustness analysis is a crucial methodological strategy because their models are often based on idealisations and abstractions, and it is usually difficult to tell which idealisations are truly harmful."* (p.541)

In their defence of the epistemic use of robustness analysis, Kuorikoski et al. then go on to both respond to charges of lack of independence of models and lack of exhaustiveness of assumptions tried. First, they try to defend the idea that we need models to be independent, albeit only independent in a restricted sense in order for robustness analysis to be successful:

> *"For derivational robustness to count as a form of triangulation via independent means of determination, the different derivations of the same result should be somehow independent."* (p.542)

At the same time, they want to defend the claim that we do not need the assumptions tried to be exhaustive in order to lose at least some of our worries about unrealistic assumptions. We will look at their arguments in more detail in the next two chapters, but for now, the important thing is that only one of these defence strategies would be enough for them to defend the confirmatory value of robustness, since they each defend the soundness of different kinds of arguments.

Making the distinction is even more important when criticising the idea that robustness analysis can offer confirmation: Showing that one argument does not work does not need to mean that the other one does not, so critics need to attack both. As we will see, critics have typically concentrated on finding fault with the idea that models can be independent sources of evidence, thereby leaving much of the literature unscathed.

So there is use in distinguishing these approaches, because it clarifies what may be needed for robustness analysis to be successful in increasing confidence in inferences from models. It seems that here we have an instance of where our taxonomy can help to throw some light on a philosophical discussion, where previously notions with different normative credentials had not been distinguished – incidentally, this seems to be exactly the kind of case Woodward has in mind, but did not provide an example of.

4.6. Summary

This chapter has examined the literature on robustness analysis in modelling and presented two arguments for why we might think robustness analysis to be confirmatory. It has thereby set up the targets for the following to chapters, which will show both arguments to be inadequate.

What we have shown in this chapter is that, contrary to what Woodward says about derivational robustness – robustness in theoretical modelling – being something altogether different from the other two kinds of methodological robustness, in fact the most influential literature on robustness in modelling has seen it as either a kind of agreement of a variety of evidence, or as a kind of inferential robustness.

Since agreement of a variety of evidence and inferential robustness may confer confirmation for different reasons it is important to distinguish between these two approaches to robustness in modelling. On the first approach, models are seen as independent sources of evidence for a hypothesis. Here the challenge is to show that the models compared in a robustness analysis are indeed independent sources of evidence: That they do not share the same biases, while at the same time all having some degree of reliability when it comes to making the inference of interest. On the second approach, robustness analysis may dispense worries about assumptions we are uncertain about, such as 'unrealistic assumptions': It aims to show us that certain problematic assumptions are unimportant for determining a result, and hence make us more confident in inferences from a model. In doing so, it relies on the exhaustiveness of alternative models tried.

In the next two chapters, I will argue that the two approaches to robustness in modelling described in the last chapter fail to establish that there is confirmatory value to robustness analysis in economic modelling – the practice that we identified with Banerjee's model in chapter 2. Proponents of the accounts, such as Kuorikoski et al., want to say not only that robustness analysis has the potential to be confirmatory, but also that robustness analysis, as it is in fact conducted, usually has some confirmatory value. I would like to argue that practice does not in fact add confirmatory value, and that we ought not try and change it in the way the accounts may envisage.

5. Against Robustness Analysis as Seeking Agreement of a Variety of Evidence

This chapter will criticise conceiving of robustness analysis as establishing the agreement of a variety of evidence, and the next will deal with robustness analysis as inferential robustness. As we will see, the main problem for the first account here analysed is that for it to be successful, the different models compared in a robustness analysis have to be *independent*: They cannot share similar biases, and cannot have been selected for according to whether they produce the result of interest. But in actual economic practice this is not guaranteed. In order to make this argument, let me first introduce some more flesh to my case study on models of herd behaviour.

5.1. Banerjee's Tail: Herd Behaviour and Informational Cascades

Let us first summarise some of the results from what we said about Banerjee's model above. Banerjee's model is designed to explain the phenomenon of herd behaviour – everybody making the same choice, even when their private information tells them to do something different – with appeal to imperfect information. The scenarios he has in mind range from restaurant choice to financial markets. Essentially, what he describes is that people take other people's choices as an indicator for their information. This may make people disregard their own information. As a result, information is not used optimally, and everybody may end up in the same place, even if that is the wrong choice by everybody's standards.

He then presents a formal model that makes a whole host of very specific and largely unrealistic assumptions, such as that all agents have Bayesian rationality and follow certain tie breaking rules. In a section on extensions, he then offers a number of modifications of the model, under which the main results of the model remain the same. Let us list these again:

1) If the first agent to choose i* gets a higher reward than all subsequent agents, the qualitative results can be shown to remain, but herding is somewhat mitigated. Banerjee further hypothesises that in general, decreasing rewards will tend to mitigate herding, and increasing rewards will tend to increase it.
2) We can relax the informational requirements somewhat: The result remains the same when agents do not know the order of the previous choices, but only the distribution.
3) If agents can choose to wait, but waiting is costly, the results are similar.
4) If signals can be obtained by agents at a cost, there will probably be even more herding.
5) According to preliminary examinations, results are similar for a large but finite number of options.
6) The results also remain similar when there are a small number of different types of signals.

We can now note a few things about these extensions before asking whether our two approaches to robustness can make sense of what Banerjee is doing here. Firstly, note that only very few of the assumptions of the original model are modified. Notable exceptions are the assumptions of Bayesian rationality, the choice of solution concept and the independence of signals. Consequently, the modifications that are made are made against the background of these assumptions that are not challenged. Secondly, they are also made each separately, and not in combination. Thirdly, for several of the changes, for instance (4), only one alternative is tried.

It may be unfair to only look at one paper and take that to be exemplary of the practice of robustness analysis. When Kuorikoski et al. (2010) claim that economic modelling practice essentially *is* robustness analysis, they did not mean that this robustness analysis is necessarily carried out by one single economist. They think that robustness analysis can be seen as a collaborative project.

> *"The modified models are often, although not necessarily, presented by different economists than the one(s) who proposed the original model (as shown below, this holds for our case study). In this sense, then, our claim is that theoretical model building in economics is to be understood as collective derivational robustness analysis."* (p.549)

And indeed, we often find that economic theorists respond to some model by developing modified versions of that model, so that we find a whole family of similar models. For instance, a closer look at the literature reveals that Banerjee's model of herd behaviour is just one of the most influential models of herd behaviour, and one that inspired many followers.[5]

Hirshleifer and Teoh (2003) provide a survey and a taxonomy of models of herd behaviour. On their definitions, Banerjee's model describes a case of an 'informational cascade': A similarity in behaviour which is brought about by agents' choices being affected by the observation of other people's actions, where agents start to disregard their own private information. Hirshleifer and Teoh take a model very similar to Banerjee's as a baseline case and then present a variety of different models that share the core features of informational cascades. They take this to show that the specific features of the baseline model are not necessary for informational cascades to occur. In particular, informational cascades can occur when…

> … options are discrete, or indeed binary (Bikhchandani et al. 1992)
> … options are continuous but bounded (Chari and Kehoe 2000)
> … there are investigation costs (Burguet and Vives 2000)
> … only a statistical summary of past choices is observed (Bikhchandani et al. 1992)
> … past choices are observed with noise (Vives 1993 and Cao and Hirshleifer 2000)
> … agents can choose to delay choice (Chamley and Gale 1994)

[5] The case study from geographical economics that Kuorikoski et al. use also involves looking at follow-up models to a baseline model that were constructed by different economists. As already announced in chapter 1, I provide a discussion of their case study in the appendix, to dispel worries that my conclusions depend on my own choice of case study. Their case shares all the problems that this chapter and the next will point to.

… payoffs are observed (Cao and Hirshleifer 2000)
… agents are imperfectly rational and use rules of thumb (Ellison and Fudenberg 1993)
… signals are public (Bikhchandani et al. 1992)

Hirshleifer and Teoh, too, use the language of robustness in their survey: When these modellers come up with different models to derive the same result, they demonstrate the robustness of that result. What is different in this collaborative project compared to the robustness analysis we find within Banerjee's paper, is that the models compared here are often different from each other in a variety of ways – it is not the case that just one assumption is varied.

For instance, in Chamley and Gale's model, that agents can delay their choice is not the only thing that is different from the baseline model in Hirshleifer and Teoh, or from Banerjee's model. In Chamley and Gale, choices are also binary in the sense that agents can either invest or not, although the time to invest can be freely chosen. Further, there are two types of agents – some who get the opportunity to invest and some who don't. The payoff from investment is uncertain, but has positive expected value, and is positively related to the number of agents who have the opportunity to invest. Whether one has the opportunity to invest is private information. Now it may be profitable for agents to wait to invest in order to gather more information about the number of agents who had the opportunity to invest. But, because waiting to invest means withholding information about one's own opportunity to invest, the possibility of strategic delay dissipates all the potential positive effects of being able to observe the market and learn. As a result, in this model, equilibrium profits are the same as if nobody could observe anybody else's actions. Everybody waits, and decisions tend to all be made very quickly, after the first agent invests. This has been interpreted as a kind of informational cascade. But clearly, this model is very different in many respects from Banerjee's.

Cao and Hirshleifer's model, too, does not only differ from Banerjee's in that past payoffs can be observed. Here choices are also binary between two projects, and payoffs from choices are uncertain in the following way: There are two possible payoffs, and projects can have two states. The state is uncertain, and for each state, payoffs are stochastic: The states differ in the probability with which each payoff is realised. Signals do not concern the payoffs, but the state of the projects. Again, an informational cascade occurs, but the model differs not only in the fact that past payoffs are observed, but also has an informational structure that is quite different from that in Banerjee's model.

The important thing to note for now is that once we go to the level of comparing models that yield similar results, but were constructed by different authors, we often find that models differ in a variety of ways, not just individual assumptions. A second important thing to note is that despite these differences between models, there are important similarities between models that are hardly ever challenged. For instance, almost all the models discussed by Hirshleifer and Teoh use the Perfect Bayesian Equilibrium solution concept, where agents are perfectly rational, there is common knowledge of rationality and setup structure, and beliefs are formed in accordance with Bayes' Rule.

If our findings from this case study generalise, then we can state the following observations, which will prove relevant for the discussion of the two approaches to the confirmatory value of robustness analysis in modelling:

When robustness is discussed in the context of a particular model, then

a) only few assumptions are altered, while all other assumptions are held fixed.
b) assumptions are altered each individually, and not in combination.
c) often, only one, or a small number of alternatives is tried.

When different authors have come up with alternative models to derive the same result, then

d) Models tend to differ in a variety of ways, and sometimes do not have very much in common apart from the herding result.
e) Nevertheless, some assumptions are rarely given up, such as assumptions of rationality, common knowledge of rationality and game structure, and the use of Nash-type equilibrium solution concepts.

Let us now see what these observations mean for the two approaches to robustness analysis in modelling that we identified above, starting with the first, which sees the models compared in robustness analysis as a variety of pieces of evidence supporting the same hypothesis.

5.2. Problems with the Account

While most of the literature on robustness analysis in modelling that we looked at in the last chapter sees robustness analysis more like inferential robustness, we have also seen some talk of robustness analysis as attempting to provide a variety of independent evidence. In fact, most of the critics of the confirmatory value of robustness analysis in modelling have seen robustness analysis as trying to furnish independent sources of evidence, as in the first approach identified above, rather than as inferential robustness: the most common argument levelled against the confirmatory value of robustness analysis claims that there is a lack of independence between models – which is a criticism which bites only against this account. Hence, there is no shortage of arguments against this idea of robustness applied to modelling. Much less can be found in the existing literature on modelling robustness as inferential robustness, as we will see in the next chapter.

When trying to conceive of the models compared in a robustness analysis as providing a variety of evidence for a hypothesis, what complicates things is that it is controversial how exactly, or even whether we can learn from the kind of mathematical models we find in economics at all in the presence of an abundance of idealisations. Given this 'gap' between model system and real world target, which is often said to be particularly large in economics, some have been sceptical of the very idea of trying to triangulate empirical results using models. After all, all we would be doing is comparing models with models: All the while we would be speaking about artificial systems, alternative ways we made up to think about a certain target (see Guala and Salanti (2002) and Sugden (2000) for these fundamental worries).

As we said above, outright scepticism about our ability to make inferences from models is probably inappropriate and in any case not widespread amongst methodologists. However, what we can say is that the fundamental doubts about what licenses these inferences mean that we can usually not quantify our confidence in our inferences from models, and that there is a deeper kind of uncertainty with regard to them. This makes it difficult to apply the kind of Bayesian variety of evidence arguments presented above.

But even granted that variety of evidence arguments can in principle apply in the case of theoretical models, more specific arguments have been made against viewing robustness analysis as providing a variety of evidence. Two arguments in particular are often levelled against robustness analysis in modelling: An argument from lack of independence, and an argument from model design. Both of these arguments contest the idea that when models agree, the best explanation lies in the truth of the hypothesis these models support, which is the normative basis for the confirmatory value of this kind of robustness. The first argument claims that agreement of models can be explained by their being so similar and hence sharing biases, and the second claims that agreement can be explained by models being designed to yield some result.

5.2.1. Lack of Independence

It is often claimed that the models compared in robustness analysis in economics do not have the necessary independence to jointly increase our confidence in a hypothesis by much. Stated like this, this is a descriptive thesis: modelling practice does not live up to the standard of independence. Still, critics also seem pessimistic about the possibility of this standard being met, and eager to point out how demanding this standard is, which points toward the normative claim that robustness analysis in modelling could not possibly be confirmatory in the way this account envisages. Lack of independence has been pointed out, for instance, by Orzack and Sober (1993) and by Cartwright (1991).

Above, we said that the necessary kind of independence for variety of evidence arguments is that having observed one piece of evidence does not make observing the other piece of evidence more likely, conditional on the hypothesis being true. In the case of measurement robustness, where as we said, sometimes authors have been concerned with deeper uncertainty about the reliability of measurement instruments, it is less clear what could be meant by 'independence'. Informally, the requirement is that two different measurement procedures do not share the same biases. But if we are uncertain about what kind of biases they may have, all we can go by is whether they are materially different, perhaps because they make use of different kinds of causal mechanisms to measure a quantity. This has sometimes been called "ontic independence" (Stegenga 2009), and may give us some confidence that two measurement procedures are not biased in the same way.

So it is to some extent unclear in the literature on measurement robustness and variety of evidence what exactly is meant by 'independence'. But what critics claim is that the models that are compared in robustness analysis are not independent on any notion of independence, because they typically share many assumptions. As we can in fact see in Banerjee's model as well as the other model variations on informational cascades in the literature, often what is

done is that single assumptions, or a small number of assumptions are varied while keeping large parts of the model fixed. The different models are then to a large extent 'the same thing' and hence we cannot speak of ontic independence – the differences between the models cannot dispense worries that they may share the same biases. Models sharing many assumptions makes it unlikely that models, if we do take them to be evidence with a quantifiable reliability, are independent pieces of evidence in the sense that the likelihoods of the evidence are independent. For instance, one assumption that all the model variations Banerjee looks at share is that private signals are independent of each other. If we are dealing with a situation in which this assumption is dubious, because this is not strictly speaking true in the target, then it may just be that none of the models indicates correctly, and all for the same reason – they may all share the same bias because they share that assumption. Even when we compare models by different authors, as we said, some assumptions are hardly ever given up, such as rationality assumptions and Nash solution concepts. In that case, we cannot dispense with the worry that it is these assumptions that are responsible for the agreement between the models.

What critics argue is that lack of independence between models means that their agreeing is not the kind of coincidence that is epistemically useful. Measurement robustness and variety of evidence get their epistemic force from the fact that the best explanation of an agreement between different sources of evidence is that the hypothesis is true. But if many assumptions, or even just a small but crucial number, are shared between models, then there is another plausible explanation, namely that the models point in the same direction because of what they have in common.

As we noted above, Kuorikoski et al. (2010) also sometimes speak as if they were defending this first approach of model robustness as the agreement of a variety of evidence. In particular, they spend an entire section of their paper devoted to independence, where they recognise that the sharing of many assumptions means that models in their entirety are not independent from each other. They maintain however that

> *"[i]t is important to realise that even though the various models […] are not independent because they share some assumptions, it is the* independence of individual tractability assumptions within a set of similar models *that is crucial for derivational robustness, rather than the independence of models"* (p. 559)

This claim is not further explained and indeed it is hard to understand what could be meant here. Firstly, it is by no means obvious what "independence of individual assumptions" could be. Insofar as they are different from each other, the alternative assumptions tried typically imply each other's falsity: for instance 'agents choose between two options' implies the falsity of 'agents have a continuum of options'. So our belief in the truth of one assumption is not independent of our belief in the truth of another. What could be meant is something like, conditional on the hypothesis being true, our confidence in the 'adequacy' of one assumption is independent of our confidence in the adequacy of another – where adequacy indicates whether an assumption is conducive to the model as a whole indicating correctly. As we will

see later, whether an assumption is adequate typically depends on the rest of the assumptions made, which again makes it odd to speak of the independence of individual assumptions.

Even if it were possible to make sense of the notion of independence of individual assumptions, it is not clear how it should be useful. If we are taking whole models to be evidence for some hypothesis, then what counts is the independence of the whole models: biases could be introduced by what all the models share, even if the assumptions in which the models differ are 'independent' - and that would take the force out of a variety of evidence argument. If the assumption of Bayesian rationality introduces a bias, then it does not help much if the assumptions that are varied – say, concerning the number of options – are independent from each other. The shared bias may still be the best explanation for agreement of the results, rather than the correctness of the result. Things are of course different if what we think is confirmed is everything that the models share: All the assumptions apart from those that are varied. The problem here is that what all the models share typically also includes some assumptions we know to be idealisations, assumptions we are in principle also worried about when it comes to the empirical adequacy of a model. And we are not interested in confirming conditional statements including those idealisations: we cannot wish to confirm something we know to be false.

The most plausible interpretation is that what Kuorikoski et al. have in mind is really the second approach identified above, which is more akin to inferential robustness and in which aims at showing the irrelevance of individual assumptions. But, as we have said, what counts there is not so much independence but exhaustiveness of the options tried. Hence, Kuorikoski et al.'s attempts to re-establish a notion of independence seem futile.

Before we accept the argument from lack of independence too quickly, however, let us remember that it is not always the case that alternative models that are said to establish a robust result share many assumptions. One of our observations from looking at models of informational cascades constructed by different authors was that the models compared here are typically very different from each other. If we disregard the fact that they typically do share some assumptions, in this case we do not have to be worried that the agreement of models is best explained by similarities between the models. Unfortunately, however, in these cases, the second argument mentioned above - an argument from design - tends to bite.

5.2.2. Model Design and Lack of Experimental Character

Odenbaugh and Alexandrova (2011) make the observation that simple economic models could be designed to yield a certain result: One assumption is varied, and a number of other assumptions are tweaked so as to preserve the result. In that case, too, it is no surprising coincidence when a number of models agree on the particular result that we tried to preserve. In this case, the models point in the same direction not because the hypothesis is true, and not because they share many assumptions, but because they were intended to point in the same direction. We already knew before that they would yield a certain result, hence we cannot be surprised by the result. In this case robustness analysis lacks experimental character: Assumptions are not changed to see what happens to a result. In a sense, the result is held fixed, and the modeller explores what sets of assumptions still yield that result. The analogy to

methods of measurement would be to only use methods of which you already know that they will yield a certain result. To conclude from these methods agreeing that there was a confirmatory boost would be cheating. Call this the argument from model design.

That models are designed to yield some result is exactly what seems to be going on in many of the models of informational cascades that we looked at above. What many of these models aim to show is that informational cascades are *possible* in a model which has property x (public signals, delay of action etc.). But it is not the case that property x is the only thing that distinguishes the model from the baseline model. In fact several other aspects of the model are changed as well. And these additional changes are made in order to preserve the cascade result, not for instance, to study the robustness of the model to changes in an ensemble of assumptions.

For example, keeping Banerjee's model in mind, it may seem quite counterintuitive that informational cascades can occur when agents can observe the payoffs to previous agents, as in Cao and Hirshleifer (2000) – Banerjee's model seems to be all about the consequences of actions being unknown. A closer look at Cao and Hirshleifer (2000) reveals that their model differs from Banerjee's in several other crucial respects. Most importantly, in their model, outcomes are uncertain and dependent on a state variable, and signals are about the states of the world. The possibility of informational cascades derives from this particular kind of uncertainty. In fact, Cao and Hirshleifer themselves report that in most models of social learning where payoffs to others are observed, informational cascades cannot occur. What they want to show is that under some conditions, informational cascades are still *possible*. Much the same holds in the case of Bikhchandani et al. (1992) showing that informational cascades are possible, and can in fact get worse when signals are publicly disclosed. Intuitively, it would seem that public disclosure would mess up Banerjee's results – it is only in the particulars of Bikhchandani et al.'s model that this result holds.

In these cases, not only have the models been designed to yield a result, and hence their agreeing with other models is not an epistemically useful coincidence. But we also know, or can be reasonably confident that had we not tweaked some other assumptions, we would not have gotten the result we wanted. So in fact we know we do not have a result that would be robust when certain other alterations were made to the model. Because models have been designed to yield a certain result, for all we know the outcome only occurs under each of the specific sets of assumptions of the models we have. This should be very troubling to those who want to argue that we can find robustness in the collaborative accumulation of models by different authors, and that this robustness can be confirmatory. It also suggests that the purpose of the accumulation of models of informational cascades is quite a different one, as will be explored in the last chapter.

Even though we often know that had only some of the changes been made, the result would not have been preserved, as in the case of allowing for the observation of past payoffs, these results are not typically reported, or in any case not in order to demonstrate the failure of robustness. Models tend to be more interesting, and more likely to be published, if they yield some interesting result. So either we want them to in fact be a model of an informational

cascade, or yield some other interesting result. This means that there is probably some bias towards models that show the robustness of some result and against models that would show the non-robustness of that result – unless the latter model is interesting for some other reason. If robustness results are more likely to be reported than non-robustness results, then again we have an alternative explanation for the agreement of models: They have been selected according to their producing the same result.

It may be objected that survey papers like that by Hirshleifer and Teoh select according to the explanandum - herd behaviour - and hence it is natural that they only report 'robust' results. Of course they will not report the cases of failures of robustness in such a survey. But firstly, a search through the articles citing Banerjee's original model does not reveal an article that evidently contests the robustness of his results, unless that article is presented as a model of something else. And secondly, it is quite telling that it seems implausible that anybody would write a survey that would report in which cases we do not get a herding result. This calls into question whether it makes sense to call robustness analysis a collective endeavour – nobody seems to keep track of the failings of robustness. What is collected are always "models of…" herd behaviour, or some other phenomenon. This suggests that economists are not in fact interested in establishing robustness in order to confirm some result, at least not in the process of the collective accumulation of models.

Putting the collective practice aside, we can probably speculate that selection takes place even in the robustness analysis we find in the context of individual models, such as at the end of Banerjee's paper. When one can only present a limited number of extensions, and one wants to 'sell' one's model, then one will only present those extensions for which one can show one's main results to be robust, not those for which one cannot do so.

The selective reporting of robust results, and deliberate design to preserve some result appear to be ineliminable from economic practice. But when there is selection according to the result whose robustness we are interested in, we cannot argue for the confirmatory power of robustness analysis using variety of evidence arguments or appealing to measurement robustness: There is a better explanation for agreement than the truth of the hypothesis the models support.

5.3. Descriptive and Normative Failings

Jointly, the arguments from lack of independence and from model design make a strong case that appeals to variety of evidence cannot show that there is confirmatory value to practices that could be or are described as robustness analysis in actual economic practice. Still, it could be objected that this is merely a descriptive claim: By a descriptive argument, what I have in mind is the claim that current practice in economic modelling – roughly the practice we identified with Banerjee's model - does not fulfil the standards of the two accounts of robustness. This causes problems for Kuorikoski et al. insofar as they want to justify current economic practice. But such an argument may not be regarded as strong or philosophically interesting when it leaves open the possibility to change the practice so that it can fulfil these

standards – we may call for a normative argument against the confirmatory value of robustness analysis.

A normative claim would say that there is no point in attempting to get confirmation out of robustness analysis – either because there is no hope for the practice ever to fulfil the standards of the accounts we looked at, and that hence we should not try to make the practice fit the standards, or because the accounts fail to show that robustness analysis, even if conducted as the account requires it, would be epistemically useful.

Still, normative arguments of this kind are often intimately connected with descriptive claims like the one developed in this chapter. And indeed, the descriptive and normative claims are often intertwined in the debate, and authors are seldom explicit about which kind of claim they want to make. It is probably safest to assume that both sides want to defend both kinds of claims. In general, I can see three ways in which this may be the case. Firstly, it can be part both of a descriptive and a normative argument to show that the accounts are more demanding than commonly acknowledged by their proponents. Secondly, the fact that a practice does not meet the standards of the account (the descriptive claim) may be evidence for the claim that it cannot do so. Thirdly, there must be some reason why the practice does persist, even if it cannot be justified by one of the accounts described. This may be evidence for the claim that modellers have a different purpose in mind – in which case it may be better to improve the practice by the standards of this purpose, rather than make it fit an account that sees confirmatory value in robustness analysis.

What I want to claim here is that what has been presented in this chapter also makes the normative claim that we should not try to aim for confirmation using robustness analysis in this account envisages plausible. Firstly, we have seen that independence is a very demanding standard, and that it is not clear what it would mean for two models to be independent pieces of evidence. We would have to use alternative models, each of which we do find to be plausible pieces of evidence, but which we are fairly confident share none of the same biases. But there are certain entrenched practices of modelling which mean that certain assumptions, such as the ones concerning rationality and equilibrium concepts, and the mathematical representation of options are rarely given up. This does not mean that modellers are not worried about them introducing a bias, and hence we can never be sure if models agree because of a shared bias. But there simply may not be the theoretical resources to come up with models that we are sure do not share the same biases and that we also think are plausible devices for making inferences about a target. So there is little hope for economic practice to ever inspire confidence that models are independent in the relevant sense.

Secondly, the case studies we looked at suggest that for most economic models, were robustness analysis carried out more systematically, we would find that most results are not robust, in which case we have learnt little from our robustness analysis. If we were to correct for the selective reporting and construction of models that show the robustness of a result, robustness would probably be a much less widely spread phenomenon – in this case confirmation through robustness analysis may not be a very effective strategy of confirmation.

And finally, especially the argument from model design suggests that economic modellers are really after something quite different when they derive the same results with different models, in which case we should investigate whether this alternative purpose is not a more worthwhile endeavour. The last chapter argues that it is.

I find these arguments to be sufficient to reject the first approach to arguing for the confirmatory value of robustness analysis in modelling. But even if robustness analysis in modelling cannot be viewed as a kind of measurement robustness or as furnishing a variety of evidence, and if there is no hope for arguing for the confirmatory value of robustness analysis along these lines, we may still think that robustness analysis as inferential robustness may be more successful as an approach to robustness analysis. After all, this is the approach taken by most of the authors who believe there to be confirmatory value in robustness analysis, such as Levins, Weisberg, and for the most part Kuorikoski et al. and Wimsatt. The next chapter will offer arguments that it is not. While showing that it is not requires different arguments, interestingly, the reasons for the failure of this account come down to the same features of modelling practice that caused problems for the first account.

6. Against Robustness Analysis as Inferential Robustness

This chapter will point to problems with the second approach to arguing for the confirmatory value of robustness analysis identified in chapter 4, namely robustness analysis as inferential robustness. After having rejected the first argument for the confirmatory value of robustness analysis for reasons of lack of independence in the last chapter, we will argue that this approach does not do any better than the first approach in justifying actual economic practice, or offering a perspective in which robustness analysis could potentially be confirmatory. Only in this case, the problems are connected in one way or another with a lack of *exhaustiveness*: a failure of robustness analysis to look at an exhaustive list of alternative models.

While the argument discussed in the last chapter has received a good deal of criticism already, the argument this chapter is dealing with has received relatively little attention. Seeing that this argument, which sees robustness analysis as a form of inferential robustness, is in fact more prominent amongst the proponents of the confirmatory value of robustness analysis, this is a surprising state of affairs - which may stem from the failure to distinguish the two arguments. This chapter hence tries to fill what may be described as a gap in the literature.

To recapitulate, the basic intuition behind seeing robustness analysis as aiming for inferential robustness was that we are worried about inferences we are making from models, and some of these worries are due to specific assumptions we find problematic. By trying out a variety of alternative assumptions, we may be able to show that the modelling result we are interested in does not depend on the problematic assumption, and hence gain confidence in an inference. Consider again the way we put the basic argument in the last chapter: We want to use model M to make an inference about a target system. We are particularly interested in one result of the model, R. Yet, we are worried about our ability to use the model for its purpose because of some unrealistic assumptions U1…N.

1) If we find out that one of the assumptions we are worried about Ui is unimportant for the derivation of R, then we gain confidence in our inferences from M.
2) Robustness can teach us that an assumption Ui is unimportant for the derivation of R.
3) Hence robustness can increase our confidence in our inferences from M.

There is clearly something to this argument. For instance, Banerjee's claim that the substantive results of the model stay the same when considering a large but finite number of options instead of an infinite number seems reassuring. We feel that we can now rule out that the result is just a mathematical oddity, as they sometimes arise when we deal with infinity.

In addition, the argument also makes sense of the fact that robustness does not help us much when it does not concern 'worrisome' assumptions. And this shows an important qualification to the simple claim that robust theorems are more useful or interesting than any other aspects of a model. Consider the following example: In the 1970s and 1980s a variety of models have been constructed in the economics of industrial organization to explain the phenomenon of

limit pricing, which occurs when monopolists lower prices to drive competitors out of the market, or deter them from entering. Once game theory had entered the toolkit of economists, several explanations were developed. For instance, in Spence (1977) excess capacity as a commitment mechanism for entry deterrence played a central role, Milgrom and Roberts (1981) emphasised reputation building, and Milgrom and Roberts (1982) accorded it a signalling function. All of these models yield a limit pricing result – in fact they were designed to do so. Yet, these models differ in their substantive assumptions: they essentially describe different mechanisms. And everybody involved in the debate saw these models as competitors. Still, they share many tractability assumptions, for instance the assumption that firms are perfect profit maximisers, an assumption which is in fact shared by most models in industrial organisation. But nobody would formulate a robust theorem that associates profit maximisation with limit pricing, or in any case it is not clear what such a robust theorem should tell us: we know that the tractability assumptions describe nothing within the target system. So if a robust theorem tells us that they are in fact responsible for a result of interest, we can say neither that we have confirmed that the tractability assumptions describe what is responsible for an explanandum, nor that the model result is confirmed, nor that the truth of the tractability assumptions are confirmed - since we know they are not true in the target.

We can say, then, that robust theorems are only potentially interesting when they associate substantive assumptions with a result of interest. The above argument gets this right. Another way of expressing this, using the idea of inferential robustness, would be that we can only view the process of finding out about the irrelevance of some assumptions as improving an inference from 'data' to a target, if the assumptions describing the data, i.e. the substantive assumptions, are indeed what is held constant between the models, and it is the auxiliary assumptions, or tractability assumptions that are varied.

Still, there are problems with each of the premises of the argument just given. This chapter will argue that we are usually not licensed to conclude from robustness analysis that particular assumptions are unimportant for the derivation of a result. Further, even if we did find this out, this does not generally license increased confidence in inferences from the model. Both of these claims have to do with a kind of non-exhaustiveness of the alternatives that are tried out in robustness analysis: All too often, too few alternatives are tried for a worrisome assumption, or they are tried against the background of keeping some assumptions fixed, i.e. not altering all the worrisome assumptions. Another main problem with the argument just given is that whether an assumption is relevant, or whether its being 'unrealistic' is cause for worry can depend on the rest of the model. Accounts of robustness analysis as inferential robustness rely on an untenable view of models as deriving their credibility for the purpose of inference from the degree of truthfulness and relative relevance of all its assumptions looked at in isolation.

6.1. From Robustness to Irrelevance of Assumptions

The second premise in the above argument claims robustness can give us evidence that an assumption is irrelevant. This section aims to show that this is not typically the case.

Let us remember again what Banerjee did when he looked for robust results in his model. He would replace one of the assumptions in his model with a different one, such as in the second extension: The assumption that agents know the order of the previous agents' choices is replaced by the assumption that agents only know the distribution of all previous choices. In some other cases, he would try out a number of alternative assumptions, such as in extension 5 – which is not formally presented, but in which the assumption that there is an infinite number of options is replaced by the assumption that there is a large but finite number of options. The latter is a composite, since there are many "large but finite" numbers – Banerjee claims that the result obtains for all of these. Still, there are also many numbers that are not "large but finite", so Banerjee had not tried out an exhaustive list of alternative assumptions to derive his result. Lastly, as we noted, the changes are made against a background of keeping other assumptions fixed. This, I argue, means that we cannot conclude from the kind of robustness Banerjee finds to the irrelevance of an assumption for a result.

In general, imagine we have a model that has a number of substantive assumptions describing some causal mechanism we hope explains some phenomenon of interest. Apart from these assumptions, the model contains a number of idealisations and mathematical tractability assumptions that do not represent the target accurately. Now assume we take one tractability assumption, A1 and replace it with a different one, A2. If we now find that the model outcome we are interested in is unchanged, or at least qualitatively the same, what have we learnt? It seems that directly, we have only learnt that *against the background of holding all the other assumptions fixed*, it does not matter that we use A1 *rather than A2*.

The problem now is that this seems to be a long step away from saying that A1 is irrelevant to the outcome full stop. Weisberg (2006b) notes that

> *"it is important to collect a sufficiently diverse set of models so that the discovery of a robust property does not depend in an arbitrary way on the set of models analysed."* (p.737)

But what does it mean for a set to be sufficiently diverse? Remember that Orzack and Sober (1993) argued that we can only be sure to get truth out of model robustness analysis when we try out an exhaustive list of alternative assumptions (i.e. one that certainly contains the true assumption), for all the assumptions we do not know to be true of the target. This way we can be sure the "true" model is amongst the ones tested, and we know the robust theorem to be true. Now we do not want to get at truth, but at a sense of relevance/importance of individual assumptions. But similar considerations apply in the case of relevance: Unless we try out an exhaustive list of alternative assumptions for each worrisome assumption, we cannot be sure that it is not relevant that we pick an assumption out of the group of assumptions that we tried, rather than a different one yet.

Above we noted that Woodward's inferential robustness gets its force from the exhaustiveness of assumptions tried, unlike arguments from variety of evidence, which rely on independence. Here, we need to be relatively certain that the true assumption is amongst the

one's we have tried out, and an exhaustive set would contain all the assumptions we think could possibly be the correct one. This may not require trying out all 'possible' assumptions, and it may not even require us to know what the set of all possible assumptions would even amount to. But unless we have some prior idea on what the correct assumption may be that allows us to narrow down the choice, we need to try out as large a variety of assumptions as we can.

Kuorikoski et al. seem to think that we can make judgements of irrelevance of assumptions even when we are sure that all of the alternatives we have tried are false. But then we cannot use our certainty that the true assumption is amongst the ones tried as a measure of what an exhaustive list of assumptions to try would be. In this case, as already hinted at in the last chapter, we could say that instead of correctness or truth, we use a standard like 'adequacy', where adequacy means that the assumption will help to derive the correct result from the model. So we would say a list of assumptions is exhaustive if we are reasonably certain that it contains the/an appropriate one. The problem here is that, as will become clear in the following, what assumption is the right one in helping a model to yield the correct result may depend on the rest of the model – which complicates things when we are also uncertain about other features of the model. Again, the implication would be that it would be safest to try out a very large number of alternatives. If this is not done, judgements of irrelevance may not be justified. We may call this problem, when the number of alternatives tried is too small, the problem of non-exhaustiveness.

A similar problem arises when modellers change an assumption against the background of holding all the other assumptions fixed – which is another feature of the practice of robustness analysis we highlighted in the last chapter. Unless we vary all the worrisome assumptions and try out an exhaustive list of combinations of alternative assumptions, we can never be sure whether an assumption might turn out to be relevant if some of the background assumptions were changed. We may call this the problem of partiality. For instance, take Bertrand's famous model of price competition (1883). Bertrand's model, as it is presented in modern textbooks (see for instance Varian 2002) has precisely two firms, identical in their production costs and the product they produce. At the time it had been established that a large number of firms competing on prices would lead to prices at the average production cost, and hence zero profits. Bertrand shows that two firms are enough to produce this result. And in fact it can be shown that any number of firms would produce this result. We may conclude that the number of firms is irrelevant to the statement that price competition leads to zero profits. Now the problem is that Bertrand's model includes a number of other assumptions that may be unrealistic in a given context. Amongst other things, as we said, it assumes that the goods produced by the firms are identical and that firms have identical costs. Sircar and Ledvina, forthcoming, show that with asymmetric costs and product differentiation, the number of firms starts to matter again[6]. So the number of firms is not irrelevant in general, but just against the background of, amongst other things, identical costs and products.

[6] Their model has free entry into the market, and they treat the number of firms as exogenous – however it can be seen from the mathematics that the number of firms matters for the price achieved:

There may be an easy fix to the problem of partiality, namely that the robust theorem we arrive at includes all those assumptions that we held fixed when checking for robustness with regard to one assumption. The problem with this approach is that it will give us a robust theorem which is little better than the original model in its entirety: we are typically worried about more than one assumption, and so not much of our fundamental worry about unrealistic assumptions will be dispersed when we arrive at such a robust theorem. Furthermore, if we carry out a string of checks for robustness on different assumptions, each against a background of holding all other assumptions fixed, as it is usually done at the end of papers like Banerjee's, all we would get is a number of slightly different robust theorems, each including a different set of unrealistic assumptions. There is no obvious way in which we could combine these robust theorems. It seems that to do that we would have to check whether the model result of interest holds also when combinations of assumptions are varied.

What may also help both in the case of partiality and non-exhaustiveness of checks for robustness is that experienced modellers can often make inferences from the behaviour of one model to the behaviour of a whole class of other models because one can see that certain types of changes would not make a significant difference in general. Banerjee clearly does so in the robustness checks he performs – in fact he makes it explicit that some of the robust results he proclaims are only informed guesses. He uses his expertise in modelling to make these judgements. Similarly, an experienced modeller may be able to tell whether two assumptions are independent from each other in the sense that we can examine the importance of one independently of the exact specification of the other.

Still, the standards necessary to license a judgement that an assumption is truly irrelevant to a conclusion simply are extremely demanding: For each assumption we find problematic, we would have to be reasonably certain that we have tried out an exhaustive list, or one that contains the assumption that is correct or appropriate for sure (where things may be complicated by the fact that appropriateness may depend on how the rest of the model is specified), or, that if we were to try out this list, we would find a robust result. Further, we would have to try out all the possible combinations of all the possible alternatives to all the assumptions we find problematic, or, again, be reasonably certain that if we were to do so, we would find a robust result.

And indeed it is hard to think of an example where an assumption could be thought of as simply discharged. In particular, we could not say this of any assumption in Banerjee's model. For instance, we do not know whether Banerjee's robust results still occur against a background of imperfect rationality – all of his extensions are made against the background of perfect rationality. We also do not know, at least not without a further model, whether results stay the same if agents can choose to wait, waiting is costly, and at the same time, those who choose early get a higher reward from choosing i* (extensions 3 and 1 respectively in section 2.2).

For instance, the price in equilibrium, with the equilibrium number of firms, differs from the price when the market is limited to two firms.

Given all this, it appears to be highly misleading to adopt a language of robustness in which we find out that a particular assumption as such is "irrelevant" or in which we can "discharge" an assumption to be left with a robust theorem as a leaner form of the original model. Neither does this adequately describe current practice in robustness analysis, nor is this a standard economic practice could realistically meet. Yet such talk is ubiquitous in the philosophical literature on robustness, and even amongst the critics of robustness as a way to improve our learning from models (see for instance Odenbaugh and Alexandrova 2011). Even if, in some very special case, we were justified to conclude that a certain assumption is irrelevant to the derivation of a result (because we are sure to have considered an exhaustive list of assumptions and partiality is not a problem), this may not help us much, since premise 1 in the argument above is as problematic as premise 2 – as will be shown in the next section.

6.2. From Irrelevance of Assumptions to Increased Confidence

Let us now turn to premise 1 in the general argument I gave above, namely that

> *if we find out that one of the assumptions we are worried about Ui is unimportant for the derivation of R, then we gain confidence in our inferences from M.*

The problem of partiality bites again with regard to this assumption. Imagine it were the case that we could establish for one model that all of the unrealistic assumptions are irrelevant for the derivation of a result. This would amount to saying that if the remaining, substantive assumptions are true, then the result is inevitable – so given they are instantiated in the target, we can be certain the result will also occur in the target. But it is extremely unlikely that such a model should exist – and the challenge is on the side of the proponents of robustness analysis to show that it does. But what if we cannot show the irrelevance of all worrisome assumptions? Then we simply cannot be sure that the irrelevance of only one or a limited number of assumptions should mean that we should have more confidence in our model.

Take the example of Bertrand's model again: Say we want to explain that no profits are made in a particular market that has an unknown number of firms, using Bertrand's two firm model. We are somewhat worried about the unrealisticness of the assumption that there are only two firms. But we also know that the number of firms does not matter for the result of interest in the model, namely that there are no profits. However, we also have doubts that the firms produce precisely identical products at identical costs. But then learning that the number of firms is irrelevant to the result that price competition leads to zero profits does not help us. Finding out about the irrelevance of this assumption does not increase our confidence in our ability to use the model to explain the phenomenon we are interested in. Admittedly, this is a very simple example, and economists or policy-makers interested in explaining market phenomena will probably use more sophisticated models. Still, I think it illustrates an important point: Finding out about the irrelevance of one assumption does not necessarily help us.

In fact, in good modelling it is often necessary that certain unrealistic assumptions, taken on their own are 'relevant' in the sense that changing them would not preserve the result. To understand the general point, take first this little story from Cartwright's (1983) *How the Laws of Physics Lie* (p.140):

> *"Imagine that we want to stage a given historical episode. We are primarily interested in teaching a moral about the motives and behaviour of the participants. But we would also like the drama to be as realistic as possible. In general we will not be able simply to 'rerun' the episode over again, but this time on stage. The original episode would have to have a remarkable unity of time and space to make that possible. There are plenty of other constraints as well. These will force us to make first one distortion, then another to compensate. Here is a trivial example. Imagine that two of the participants had a secret conversation in the corner of the room. If the actors whisper together, the audience will not be able to hear them. So the other actors must be moved off the stage, and then back on again. But in reality everyone stayed in the same place throughout."*

If we view the drama to be like a model, where the target is the episode in actual history, then here we have a case where one idealisation in a sense necessitated another in order for the model as a whole to serve its purpose well. The fact that we put the episode on stage, because we want an audience to see and hear it, means that in order for the drama to work, we have to make another idealisation, i.e. introduce another falsity, namely that the other participants will be moved off stage, away from the two participants having a conversation. Just making the first idealisation would lead to a bad model: the story would not make any sense. We need the second idealisation to set things right again. As a consequence, something like a robustness analysis on the second assumption would not make any sense: Of course it will turn out to matter that the actors were moved off stage. We will not find robustness here, but this should not be any cause for worry: We want that assumption to matter because it sets previous idealisations right.

Cartwright claims that these kinds of interactions, where two falsehoods are better than one, and one idealisation somehow necessitates another, can frequently be found in natural science as well. If that is the case, we can simply not be sure that, for instance, a model of herd behaviour that makes both the assumptions of common knowledge of the setup structure, and the assumption of Bayesian rationality will indicate correctly with regard to the hypothesis we are interested in, while a model that gives up just one of them will not. And then we should not expect one of these assumptions to turn out to be irrelevant. Against the background of keeping the other assumptions fixed, it will turn out to be relevant. Or it may be the case that given the way Banerjee chooses to characterise the options in his model, the tie-breaking rules he introduces deliver the most accurate results about a target, even though they are unrealistic, and that with other tie-breaking rules the model would simply not work.

This is a strong argument against thinking that finding that a result is not robust to changes in individual tractability assumptions is a cause for worry, which is a claim also frequently made, for instance by Kuorikoski et al. But it also means that actually finding robustness is not so

telling: In good as well as in bad models, worrisome assumptions may or may not be relevant. We have seen that an assumption that is irrelevant against some background of assumptions actually can be crucial against a different, more realistic background, and that sometimes an assumption which has turned out to be relevant should not in fact worry us, because it works well in combination with the other assumptions of a model. So finding that a worrisome assumption does not matter should not generally increase our confidence that we are dealing with a good model. At the very least, to make such a judgement, we have to have a very good understanding already of what role an assumption serves in a model. So going from the irrelevance of a worrisome assumption to increased confidence in inferences from a model is not generally warranted.

6.3. Is there Marginal Benefit to Non-Exhaustive Robustness Analysis?

I have argued that it is a problem when robustness analysis is partial, in the sense that only few assumptions are varied against the background of keeping other assumptions fixed, and non-exhaustive, in the sense that too few alternatives are tried. This is because the partiality and non-exhaustiveness of robustness analysis mean that we may not be able to conclude that an assumption is irrelevant, and that if it is irrelevant, this is necessarily cause for confidence in a model.

Still, some of the proponents of this approach to robustness analysis argue that there is at least some marginal benefit to showing that a result is robust under a non-exhaustive set of alternative assumptions. Kuorikoski et al. maintain that non-exhaustiveness is not a problem since there is at least a marginal epistemic boost for each new possibility tried, when we are not sure which assumption would be appropriate:

> *"Allowing for a non-exhaustive set implies that we cannot be sure that the relationship is actually robust, but it does not remove the epistemic relevance of robustness analysis altogether. Its epistemic import comes in degrees."* (p. 560)

Under some ideal circumstances, they would be right. Imagine that there are ten possible specifications for a particular assumption, and we are sure that at least one of them will be adequate in that it will help yield the correct result for sure. We are not worried about any of the other assumptions, so partiality is not a problem. We think each set of assumptions is equally plausible, in which case it seems that finding out that two models yield the same result should give us more confidence in that result being correct than only one model yielding that same result.

There are three problems with this idea: The first is that as above, in most cases we will be worried about more than one assumption, so the problems with testing for robustness keeping another worrisome assumption fixed arise. The second is that when there is a very large number of possibilities for which assumptions could be used, the marginal effect of knowing that two yield the same result rather than only one doing so will be minimal. When there are 100 possible, equally plausible sets of assumptions, then knowing that two rather than only

one yield a certain result, we are now 1/50 sure that one of our models gave us the right result for sure, rather than only 1/100. We have said that the normative force of inferential robustness comes from our confidence that the set of models we tried and found to deliver the same result includes the, or a, correct/adequate one. If robustness analysis is non-exhaustive we may just not have that confidence. The third problem is that this argument does not work if modellers selectively only report alternative models under which a result is robust. But as we have claimed in the argument from model design above, this is frequently the case in theoretical modelling: models are designed to yield a certain result.

As Woodward 2006 notes, it has sometimes been claimed that there is a kind of exhaustiveness-robustness trade-off when we conduct robustness analysis: The more models we try out, the less likely it is that they will all produce the same result. To call this as a trade-off, however, is very misleading. If we already know that we would probably find that robustness would not hold anymore if we tried out more and more alternative assumptions, then this calls into question the confirmatory value of the robustness we had in the first place, at least on this account. We cannot simply collect all the models under which we get a result of interest and then stop there if we want to appeal to inferential robustness. What counts is that most of the models we find plausible point in the same direction, and to establish that we cannot conduct a robustness analysis where we actively search out models that give us the desired result.

These considerations severely call into question whether there is any marginal benefit to trying out a limited number of alternative model specifications. There would only be a confirmatory boost under a very restrictive set of conditions, which I doubt is ever satisfied in modelling practice.

6.4. A Caricature

We may perhaps draw the following caricature of the argument Wimsatt, Weisberg, and Kuorikoski et al. have in mind, when they say that robustness tells us about irrelevance of assumptions, and irrelevance can increase our confidence in inferences from the model. The caricature is meant to describe the logic that seems to be underlying the arguments we discussed. We have a model M which yields a result R of interest. We would like to say that R occurs in the target, and have confidence C(R) that it does, which somehow derives from the model: We take the model to be evidence that R occurs in the target. The model makes a number of assumptions A_i, each of which has a degree of 'worrisomeness' attached to it $W(A_i)$, as well as a degree of relevance for deriving the result of interest $R(A_i)$. Our overall confidence in the model for purposes of making a certain inference derives directly from the worrisomeness and relative relevance of the individual assumptions. In particular, the worrisomeness of each assumption is weighed by its relevance, and the overall confidence we have in a model is just the converse of the sum of the weighed worrisomeness of all the individual assumptions:

C(R) is an increasing function of - Σ $W(A_i)R(A_i)$

It does not matter here, and in any case is not made precise in the accounts we looked at, how exactly C(R) derives from the negative weighted sum of the worrisomeness of the different assumptions. What is crucial, and what I want to illustrate with this here is that these authors think that the more worrisome or unrealistic any particular assumption is, the less confidence we have in the model as a whole indicating correctly, and the less relevant any particular worrisome assumption is, the more confident we can be in an inference from the model. Such a view of models appears to be implicit in the approach to robustness analysis as inferential robustness. Given this view of models, the claim now is that robustness analysis, by teaching us about Ri, can change our overall confidence in an inference from the model, i.e., increase C(R): All we have to do is show that the relatively more troublesome assumptions are relatively less relevant.

The main problem with this conception of models is that it relies on an ordering of assumptions in terms of worrisomeness and relevance: a model is more reliable when the relatively less worrisome assumptions are relatively more important for deriving a result of interest. Just like decision theory needs an ordering of options in order to come up with a numerical measure of utility, in order to have a notion of degrees of worrisomeness or relevance, we need an ordering of assumptions according to these criteria. But in order to come up with such an ordering, we have to make isolated comparisons of worrisomeness and relevance of individual assumptions. What I have tried to show is that how worrisome and how relevant an assumption is may depend on the context of other assumptions in which it is placed: Firstly, we cannot judge the relevance of an assumption for a result in isolation of what other assumptions are. And secondly, how worried we are about an assumption may also depend on what the other assumptions are. So we cannot typically say that one assumption is more worrisome than another, or more relevant than another, simpliciter. And then the kind of argument just described cannot get off the ground.

What the arguments in this chapter have shown, then, is that this is a misguided view of models. And insofar as it underlies the arguments discussed in this chapter, it illustrates well the reasons for their failure: How well a model serves an inferential purpose does not derive in a straightforward sense from looking at all the assumptions of a model individually. We cannot look at assumptions in an isolated way to determine how worrisome they are, and how relevant they are to a particular outcome. Both of these judgements tend to depend on all the other assumptions made – at least that should be the stipulation until it has been shown that it is legitimate to look at what an isolated assumption does in a model. This makes it unrealistic that we can learn much in the piecemeal way that the account of robustness analysis as inferential robustness seems to envisage: That we can proceed by playing around with assumptions individually and go from robustness to irrelevance of individual assumptions to increased confidence in a model.

6.5. Summary: Why Robustness Analysis is not Confirmatory

This chapter has argued that robustness analysis is not, and indeed cannot be confirmatory in the way that the proponents of robustness analysis as inferential robustness aspire it to be. Again, this involves a *descriptive* claim about economic practice, as well as the claim that the practice cannot and should not be changed in order to try and make it fit this account, i.e. a *normative* claim.

Interestingly, the reasons for the descriptive failure of this account are similar to the reasons for the failure of the first approach to arguing for the confirmatory value of robustness analysis, analysed in the last chapter: They have to do with the non-exhaustiveness of the assumptions tried out in robustness analysis, and the partiality of robustness analysis, i.e. the checking for robustness against a background of keeping other assumptions fixed. While in the case of conceiving of robustness analysis as aiming at the agreement of a variety of evidence, these aspects of robustness analysis were evidence of a lack of independence between models, and of an element of design and selective reporting of results, non-exhaustiveness and partiality cause different problems here: Both non-exhaustiveness and partiality mean that it is usually not warranted to speak of assumptions being irrelevant for a certain model outcome, and to allow for them to be 'discharged' to arrive at a supposedly more credible robust theorem. And partiality means that even if we did find out that an assumption was irrelevant to a result, this may not mean that we are justified in having a higher degree of confidence in a model: The assumption may just have been irrelevant against the background of the other assumptions made, and may cause problems again when we change other parts of the model.

Furthermore, to make the normative claim, the chances for eliminating problems of non-exhaustiveness and partiality seem slim. Firstly, as already argued in the last chapter, if we were to check for robustness more exhaustively, we would probably find there are few interesting robust results left. It will probably always turn out that some problematic assumptions do make a difference when we replace them individually. But this need not be cause for worry, when as we have said, assumptions may interact in ways that require individual unrealistic assumptions to make a difference. The argument that was attacked in this chapter relies on a piecemeal view of models in which our confidence in a model derives from how relevant individual assumptions are for an outcome, and how troublesome we find them. But assumptions are only relevant or worrisome against the background of the rest of the model, which is always cause for worry when robustness analysis proceeds by playing around with individual assumptions. In particular, this is always cause for worry in the two premises of the argument that was attacked in this chapter: that we can conclude from robustness analysis that an assumption is irrelevant, and when we conclude from the irrelevance of an assumption that we can be more confident in an inference from a model.

That such a misguided view of models seems to be underlying these arguments may be a symptom of these authors wanting to avoid committing to any comprehensive theory of how we learn from models. They aim to present a way in which we can increase our confidence in inferences from models in an incremental way without putting it into a context of what makes it the case that we can learn from models in general. In our discussion, however, it emerged

that we probably cannot do so. When the role of individual assumptions cannot easily be isolated from what the model as a whole is doing, then we need a more comprehensive account of modelling to make judgements about whether we want an assumption to be relevant for a result or not, and whether robustness is a good indicator for the relevance of an assumption. By trying to avoid thinking of a more comprehensive account of why we can learn from models, these authors explicitly adopted one that implausibly sees the reliability of a model as resulting in a straightforward way from the adequacy of all its assumptions in isolation.

What we are now left with is the realisation that attempts to argue for the confirmatory value of robustness analysis have failed. Robustness analysis does not provide a solution to the problem of unrealistic assumptions in the way that the literature we discussed envisaged. So both of the open questions that motivated us in the introduction have been left unanswered: how to learn from economic models, and what the value in robustness analysis is. Still, we have learnt much along the way: We have learnt about the nature of economic models and how they are composed from their assumptions and about how economists proceed in modelling. And we have seen that robustness analysis should be understood in the context of wider debates in modelling. The next part will take up these lessons and use observations about economic practice and ideas from the wider literature on modelling to suggest some alternative interpretations of robustness analysis.

PART III: Towards an Alternative Interpretation of Robustness Analysis

Part I posed the problem of what the use of the economic practice of robustness analysis was, where robustness analysis refers to the repeated derivation of the same result using models that differ in some of their assumptions. In Woodward's terms, we were not sure what derivational robustness was really good for. Part II discussed an answer that has been given by a number of philosophers of science, namely that robustness analysis aims at a kind of confirmation. But now that we have seen that their arguments for the confirmatory value of robustness analysis fail, we are again left wondering what it is that economists aim to achieve when they conduct robustness analysis. After all, theoretical economists spend much of their time deriving the same results with alternative models. Seeing that economists usually fail to get any confirmation out of this practice, are they just wasting their time? The last part of this thesis argues that they are not.

I mentioned in the last chapter that one of the ways in which the descriptive failure of an account is connected to a normative failure is that descriptive failure can indicate that there is an alternative explanation of why scientists engage in a practice. The fact that robustness analysis fails to be confirmatory in the sense that authors like Kuorikoski et al. want it to be may suggest that economists are aiming at something quite different. In that case, there is a prima facie case not to try and make practice fit the philosophical accounts described above, but to look into what economists in fact want to achieve with robustness analysis, and, if that project is more promising, evaluate practice according to those standards.

Chapter 7 will make first steps towards such an argument. It will argue that indeed economists do not have the goal of confirmation in mind when they conduct robustness analysis. Instead, I will make some observations about the case study on herd behaviour that indicate what alternative goals economists may in fact have in mind. We will see that achieving these goals is a more promising endeavour than trying to get confirmation out of robustness analysis. While I will not provide a full-blown account of robustness analysis and engage in a defence of it, this will at least point to a more fruitful way to approach the practice of robustness analysis and suggest some interesting avenues for further investigation.

The final observations of chapter 6 show that we cannot understand the practice of robustness analysis without thinking about modelling more generally, and about how and what for models are used in practice. By trying to think about robustness out of the context of the general debate on modelling, the authors we looked at adopted an implicit view on modelling which is quite misguided. The following chapter will hence try to look at the practice of robustness analysis in the wider context of what we might want to achieve when we model a phenomenon. When we lose the narrow focus on confirmation, we will see that robustness analysis has to do both with the question of how models explain, and with our thinking about the target of investigation.

7. Alternative Interpretations of Robustness Analysis

Several features of current economic practice, some of which we have already noted, suggest that economists in fact aim for something quite different when they conduct robustness analysis. Before this chapter develops an alternative interpretation of what economists are trying to do when they conduct robustness analysis, I would like to go through three prevalent features of the practice of robustness analysis that are at odds with a focus on getting confirmation out of robustness analysis:

1) Much of what economists call robustness analysis is better understood as what philosophers of science call 'de-idealisation'.
2) The models compared in robustness analysis often differ in their substantive assumptions.
3) Robustness analysis typically lacks experimental character: models are designed to yield certain results.

7.1. Evidence that Modellers are not Aiming for Confirmation

7.1.1. De-idealisation

The first significant feature of robustness analysis as it is practised in theoretical modelling that suggests that economists are not aiming at confirming results in the way the two approaches we discussed envisage, is that economists are often after de-idealisation rather than robustness as philosophers understand it.

De-idealisation is a concept frequently talked about in the philosophical literature that is very similar to robustness. In contrast to robustness, de-idealisation is not a term that is used much by scientists themselves – it is predominantly used by philosophers to describe some scientific practices. In fact, what I want to claim here is that scientists themselves often speak of robustness when they mean de-idealisation. This can be a source of confusion, because there are some important differences between the two concepts. The philosophical arguments for the confirmatory value of robustness that we discussed are concerned with robustness strictly speaking. So if it turns out that much of economic practice is concerned with de-idealisation and not robustness, then at least the practice these philosophers want to justify is not as widespread as previously thought. I want to argue exactly that: Modellers are frequently after de-idealisation when they speak about robustness in their model.

Roughly, we speak of de-idealising a model when we take an unrealistic assumption in the model and replace it with a realistic one, by which I mean one that is true in the target. It is usually seen to be an advantage when we can show that a 'less idealised' version of the model still leads to the same result. In fact, one of the most prominent accounts of how we can learn

from idealised, or unrealistic models in general is founded on the idea of de-idealisation. On this account, in order for us to be able to learn from an idealised model, the target needs to instantiate all those assumptions which cannot be ´de-idealised´, i.e. those assumptions which we cannot weaken (make more realistic) without changing the model outcome we are interested in (see McMullin 1985).

It has often been pointed out that, especially in economics, models can usually not be de-idealised enough in order to fulfil this requirement: We cannot de-idealise all assumptions not instantiated in the target while still preserving the result of interest (see for instance Reiss 2007). But de-idealisation can be partial as well: Just one, or a handful of assumptions are made more realistic while the result of interest stays the same. This is admittedly a practice that is very widespread in economic modelling, and also frequently carried out under the heading of 'extensions and modifications'.

In some sense, this practice can be seen as a kind of robustness analysis: Some assumptions are varied while some modelling result stays stable. What is special about de-idealisation, however, is that the original assumption is replaced by one that is judged to be more 'realistic', a better fit with the target than the old assumption. Take for instance the 5th extension in Banerjee's model that we identified above:

> *According to preliminary examinations, results are similar for a large but finite number of options.*

It becomes clear in Banerjee's discussion that he thinks that the assumption of a large but finite number of options is in fact more realistic than his assumption of a continuum of options. So if he were to formally present a model with a large but finite number of options, then for him this would be a model with one less unrealistic assumption. But the assumption of a continuum of options makes the mathematics of the model much simpler. So he goes on to check whether the realistic assumption would lead to the same result in order to justify the simplification in the original model:

> *"Preliminary investigations show that the results we get for the case where there are a large but finite number of options are much more complicated but quite similar. We therefore feel justified in working with this much more tractable model which we see as an approximation to the other case"* (p. 816)

Similarly, when Ellison and Fudenberg (1993) present a model of herd behaviour where agents are imperfectly rational, they are clearly motivated by the thought that this is a more realistic assumption than that of perfect Bayesian rationality.

The most obvious way to interpret what scientists are doing when they de-idealise a model is that they simply come up with a better model than the original one. In the case where we want to use a model to explain an observed phenomenon, the fact that a result stays stable is significant only in so far as that means that the better model does in fact explain the phenomenon it is supposed to explain: We can derive the observational result from it that we hope to explain, and that we already hoped to explain with the old model.

The logic here is very similar to the argument for the value of inferential robustness described above: In the case where we actually do know what a more realistic assumption would be, a more obvious way to lose worry about an unrealistic assumption than trying out a whole variety of alternatives, is to simply replace it with a realistic one. We are then left with a model we are less worried about. Because of this parallel to inferential robustness, the focus on trying to dispense with worries about individual assumptions, much of my criticism of inferential robustness in the last chapter also applies to de-idealisation, when it is carried out partially. So, for instance, showing that the unrealism of one assumption does not matter for a result does not always license increased confidence in a model. But let us put these problems aside for the moment.

So an obvious way to interpret partial de-idealisation is to say that we replace an original model with a better model. However, if this were all there was to it, then the crucial question becomes why we should still bother with the original model. Economists do not usually speak as if they wanted to dispense with the original model when they de-idealise in the extensions and modifications section of a paper. And McMullin asks us to see de-idealisation as a way of explaining why some idealised model in fact applies to a particular situation. However what could be the reason to still use the old model if we have more realistic models that can also explain the phenomenon of interest? Perhaps, if McMullin is right, we found out that the old model applies to a situation, but this is of little interest if we have a new, better model.

There are some reasons why we might still want to think about the old model even if we have successfully de-idealised it and found it still explains the phenomenon it is supposed to. Let me discuss these here, since they will be of use below in trying to understand the practice of robustness analysis. Firstly, the original model may be mathematically simpler and hence allow an easier exposition of a problem. This is the case in the example from Banerjee's model just mentioned: It is simpler to do the maths with a continuum of options than with a large finite number. The fact that the result of interest can be derived with either model justifies our still using the old model for expository purposes: After all, we found that the fact that a particular assumption is unrealistic was irrelevant.

Simplicity is often seen to be an explanatory virtue – other things being equal, the simpler an explanation, the better (see for instance Baker 2009). This may be seen as a version of the (in)famous Occam's Razor, which commands adopting simpler theories, hypotheses, or, in this case, explanations, when there is no other way, in particular no empirical reason, to distinguish between them (also Baker 2009). As a descriptive thesis, it has been suggested by experimental work that people tend to find simpler explanations more explanatory (Lombrozo and Rutstein 2004). Coming back to our discussion, we may think that the fact that both a more and a less idealised model agree on a result of interest means that we have no way to distinguish between the two, so the principle would bite.

Let us grant that there is some explanatory virtue in having an explanation that is simpler. Still, it is plausible to think that when it comes to making inferences from a model, we trust the de-idealised model more. Think for instance about prediction in Sugden's schema.

P1. In the model, R is caused by F.

P2. F operates in the target.

Therefore, there is some reason to believe:

P3. R occurs in the target.

Here we know that some factors that are causally important in the model, F, are in fact instantiated in the target, but we do not know that a model result R is instantiated in the target. We now want the model to give us reason to believe that the model result is instantiated in the target. Simplicity would seem to be irrelevant in giving us such a reason, and if the unrealism of some further modelling assumptions is the problem, then the less idealised model gives us a better reason to believe in the inference. Once we have the de-idealised model, we also have a better reason to believe in the prediction that could be made from the original model, simply because it was the same prediction that is made by the more trustworthy model. But it is only in virtue of having the new model that we have higher confidence in the old model's prediction. The new model is doing all the work.
Seeing that here the task is prediction, it is no wonder that simplicity as an explanatory virtue does not help. But a similar argument applies to Sugden's explanatory inference scheme:

E1. In the model, R is caused by F.

E2. F operates in the target.

E3. R occurs in the target.

Therefore, there is some reason to believe:

E4. In the target, R is caused by F.

Here we do know that both the causally important features of a model, as well as the model result are instantiated in the target. The inference concerns whether the result in the target is really caused by the features that matter in the model. The thought here is that a good explanation should cite the causes of the thing we are trying to explain. Again with this inference, simplicity does not help us – we are worried about unrealistic assumptions, and whether because of them, causes operate differently in the model than they do in the target. Simplicity does not give us a reason to trust one model more than the other, whereas we may think that we can trust the de-idealised model more. Again, the new model may give us reason to believe that the causes cited in the old model operate in the target. Once we have this knowledge, we may think that simplicity means that the old model is in fact more explanatory. But it is in virtue of the new, less idealised model that we gained the extra confidence in the causal inference. So simplicity may be an explanatory virtue, but for purposes of making inferences, de-idealisation attempts to replace one model with a more trustworthy one.

A second reason why we may still be interested in the original, more idealised model, is that the old model may allow for explanatory unification. Imagine we want to explain some phenomenon such as herd behaviour. A phenomenon, being a type, is something that can occur in a variety of different circumstances. For instance, 'herd behaviour' is a phenomenon that is said to occur both in financial markets and in the case of restaurant choice. To de-idealise a model a model of herd behaviour, we do not only need to know what the type of

herd behaviour is, but we also need to know enough about the detail of the specific instance of the phenomenon we want to explain in order to have the required knowledge of the state of the target system to be able to judge what is 'realistic' relative to the target.

To take the case of Banerjee again, whether a continuum, a large but finite number of options, or just two options are 'realistic' depends on which instance of herd behaviour we are talking about. Banerjee, for instance, seems to have a situation in mind where there is a very large but finite number of options. Still, as the exposition of Banerjee's model makes clear, he also hopes that we can apply his model to all sorts of different situations. But for each of these situations, de-idealisation would mean something different. In fact, the first situation that Banerjee himself describes only includes two options: it is the choice between two restaurants. He even presents a very simple model which is similar to the one he later develops which applies to this binary choice. So a kind of de-idealisation to a two-option scenario is also possible. In fact, many of the models of herd behaviour that have been constructed after Banerjee first published his are restricted to a binary choice (as for instance Bikhchandani et al. 1992 and Chamley and Gale 1994). What we seem to have here is a situation where we could de-idealise one model in two different directions, one each for two different kinds of situations where we have observed a particular phenomenon. If we just had two different models for these two different situations, then we might miss important similarities. The 'core' model can offer a kind of unification of these two situations in which we observe a certain phenomenon. In fact, without the model, we may not even speak of the same phenomenon in both cases, because we might have missed some similarities without the bridging model.

Unification, like simplicity, is often seen to be a virtue of explanations. In fact Kitcher (1989) has developed an account of explanation that has the idea of unification at the centre of what it means to be explanatory. Still, the same caveats as with simplicity apply: Unification does not help us with making inferences – it is not an epistemic virtue. Once we have made our inferences, possibly gaining confidence from less idealised versions of the model, unification may provide an explanatory bonus. But if we are worried about unrealistic assumptions, and a de-idealised model can make us lose some of these worries relative to an original model, then the de-idealised model is simply preferable to the old one for inferential purposes. When we are concerned with justifying inferences from a model, de-idealisation could be seen as providing us with new models that are more trustworthy than the old, and in terms of inferential reliability, there is no more reason to use the old model.

What does robustness as de-idealisation have to do with the two approaches to robustness identified above? The first sees the different models under which a result is robust as independent sources of evidence that that result is instantiated in the target. In de-idealisation, there is no more reason to use the old model as a source of evidence, since the new model is judged to be more reliable, and the old one is reliable only in virtue of what it shares with the new model. We de-idealise because we strictly prefer the new model for epistemic purposes – so variety of evidence need not be the primary consideration when it comes to de-idealisation.

The second approach we identified uses robustness to find out about the irrelevance of a worrisome, unrealistic assumption. This is more similar since in the case of de-idealisation we find out that at least when it comes to the derivation of that particular result, it did not matter that we had used a particular idealisation. But in the case of inferential robustness we conclude that an unrealistic assumption does not matter because we try out a wide variety of alternative assumptions – we do not find this out by using a realistic assumption instead.

Most proponents of inferential robustness acknowledge that it would be preferable to de-idealise rather than try out a wide variety of alternatives. They reserve inferential robustness for the cases where we simply do not know which assumption would be appropriate, or where it is not clear what a realistic assumption would even be. In the case of inferential robustness, in the end we still use the original model, even though we have examined a variety of alternatives. The alternative models serve the purpose of showing that a particular assumption is irrelevant; they do not replace the original model. After all, we have no way of knowing if they are preferable to the original model in terms of realism.

In some sense, inferential robustness is just a stand-in for de-idealisation for situations where we do not know what the realistic assumption would be. In the case of de-idealisation, we come up with a model we think is better for inferential purposes. We may still use the old one for reasons of explanatoriness, and because the new model, insofar it is itself reliable, in a sense confirmed that the old model got things right with regard to the result of interest. In the case of inferential robustness, we do not know what that better model would be. But we use many models to try and confirm the original model in a similar way as in the case of de-idealisation: by showing that an unrealistic assumption was irrelevant for a particular outcome. We keep on using the original model because we have no one better model, but we have stripped it down to a robust theorem.

What this discussion is meant to show is that robustness analysis is only ever of use when de-idealisation is not possible. This further limits the extent to which the arguments appealing to inferential robustness can ever find application. Indeed, many alleged cases of robustness analysis are better thought of as cases of de-idealisation, as our examples above show. When de-idealisation is possible, for epistemic purposes, what we are doing is trying to replace one model with a better model – although caveats about partial de-idealisation apply. For explanatory purposes, however, we may still want to think about the old model: The reason for the accumulation of models has more to do with the goal of explanation than with the goal of confirming an inference.

Focus on de-idealisation is hence one way in which economist's goals do not align with the accounts of robustness we identified in the last chapters.

7.1.2. Comparing Substantively Different Models

Both when we look at Banerjee's model and at the ensembles of models of herd behaviour that were constructed in its wake, we find that the models that are compared in robustness analysis often differ in their substantive assumptions, that is, in assumptions that are relevant for the description of the causal mechanism that is supposed to explain the herding result. For

instance, in Banerjee's model, giving agents the option to wait and see, or making it a choice whether to buy a signal or not completely changes the choices agents need to make and the model then describes a different kind of interaction. This interaction may bring about a similar result and hence may potentially explain the same phenomenon, but it describes a different mechanism bringing about this herding result.

Consider again the list of models of herd behaviour that have been constructed since Banerjee's: Informational cascades can occur when…

> … options are discrete, or indeed binary (Bikhchandani et al. 1992)
> … options are continuous but bounded (Chari and Kehoe 2000)
> … there are investigation costs (Burguet and Vives 2000)
> … only a statistical summary of past choices is observed (Bikhchandani et al. 1992)
> … past choices are observed with noise (Vives 1993 and Cao and Hirshleifer 2000)
> … agents can choose to delay choice (Chamley and Gale 1994)
> … payoffs are observed (Cao and Hirshleifer 2000)
> … agents are imperfectly rational and use rules of thumb (Ellison and Fudenberg 1993)
> … signals are public (Bikhchandani et al. 1992)

These models, too, for the most part appear to describe fairly different mechanisms, and the assumptions in which they differ can be regarded as substantive assumptions. While they can all potentially explain the phenomenon of an informational cascade, they are probably better seen as describing rival causal explanations.

But if the different models describe different causal mechanisms, they would seem to find application in different kinds of situations. For instance, Chamley and Gale's (1994) model, where agents can delay action, is used to explain situations in financial markets where for a long time nothing happens while agents wait, and an informational cascade then occurs suddenly and rapidly. None of the other models would find application here. Burguet and Vives (2000), on the other hand, is a model of learning in a prediction task, which is not specifically geared to explain market phenomena. Hirshleifer and Teoh acknowledge these substantive differences when they write the following:

> *"In sum, whether information channels become quickly or only gradually clogged, and whether the blockage is complete or partial, is dependent on the economic setting; but the general conclusion that there can be long periods in which individuals herd upon poor decisions is robust."* (p. 31)

They appear to think that different models will find application in different economic settings. What they call "robust" is the fact that informational cascades can occur – that they are possible in a variety of settings. We find that a variety of models can explain informational cascades in general, but different models may be required to explain each particular occurrence.

Why is this at odds with robustness analysis playing a confirmatory role? Most importantly, it means that robustness analysis cannot be thought of as supporting two out of the three kinds

of inferences Sugden identifies, namely abduction and explanation. In the case of abduction, what is confirmed is that the substantive assumptions are true in the target – but that cannot be the case when the models that robustness analysis compares differ in their substantive assumptions. In the case of explanation, we infer that the substantive assumptions really describe what caused a phenomenon of interest. Again, in this case, we would need the models compared in a robustness analysis to agree on the substantive assumptions.

So what about the case of prediction? Technically, here it is only required that the models agree on a result that we predict will occur in the target. This is the case for all the different models of herd behaviour we discussed, if we take the result of interest to be that there is some kind of herding – similarity in behaviour even though there is diversity in the individual information received. But here robustness analysis would only make sense when we are uncertain about which substantive assumptions are true in the target. If we did know that one set was true and another was false, then we would take the model with the true substantive assumptions to be predictive, at least we hope it is, and we would not think that knowing about another model making false substantive assumptions as warranting additional certainty in the prediction.

The problem now is that if we were uncertain about the substantive factors at work in the target to the extent that we cannot decide between two models that differ as dramatically as, for instance, Chamley and Gale 1994 and Banerjee's original model, then it seems that we know so little that we are not in a position to make a reliable prediction at all. We would have to be in a position where we both do not know that the result occurs in the target, and we are uncertain about what causal factors operate in the target. This is not a situation in which anybody would be likely to use a mathematical model like the ones we looked at to make a prediction. Further, as a matter of fact, these models are rarely used to make predictions. As we noted in the introduction, these models are usually thought of as explanatory: We already know some phenomenon occurred, and we are trying to find out what caused it.

When models that agree on a result differ in their substantive assumptions, they are best thought of as providing alternative explanations for one and the same phenomenon. As we saw earlier, this is exactly what happened in the case of models of limit pricing, and is arguably what often drives the accumulation of models in theoretical economics. But in this case, there can be no confirmatory value to robustness analysis: We already know that the result they agree on occurs, and in everything else the different models are best seen as rivals. In any one case, only one of the models will best explain the occurrence of the phenomenon, or only one of them can be used to reliably predict that the phenomenon will occur.

7.1.3. Lack of Experimental Character

A third way in which economic practice reveals that economists do not even try to get at confirmation in the sense our accounts of robustness envisage is the lack of experimental character of robustness analysis that we already mentioned in chapter 5. Economists often design models in order for them to yield a certain result, even in the full knowledge that slightly different model specification would not give them the same result. All I mean here by the lack of experimental character is that the result of the model is already anticipated when

the model is being designed, and beliefs about what the result will be influence the design of the model. We have seen that this is in the way of robustness analysis serving a confirmatory purpose.

In the case of the follow-up models of informational cascades, there seems to be a focus on showing the possibility of a cascade result under a large variety of different settings. As we noted in chapter 5, what many of these models aim to show is that informational cascades are *possible* in a model which has property x (public signals, delay of action etc.). But it is not the case that property x is the only thing that distinguishes the model from the baseline model. In fact several other aspects of the model are changed as well. And these additional changes are made in order to preserve the cascade result, not for instance, to study the robustness of the model to changes in an ensemble of assumptions.

If modellers were aiming at getting confirmation out of robustness analysis, they would not design models to yield a desired outcome – robustness analysis would have to have a more experimental character. So this again is a clue that modellers have something else in mind.

7.2. An Alternative View of Robustness Analysis

What these three features of robustness analysis as it is practised show is not only that getting confirmation out of robustness analysis is not what economists are aiming for when they conduct robustness analysis. They also suggest some alternative goals that economists are in fact aiming at when they conduct robustness analysis. While I cannot offer a full defence of these goals here, I would like to elaborate them, and give them a rationale that is at least plausible and involves goals that are more easily attainable than confirmation.

What is most striking about the observations about economic practice just made is the following: Most of the time, when modellers derive the same result with different models they either want to introduce a model that is better than the original one, or they want to introduce a model that is simply different, and will apply in different situations than other models. In neither case is there any confirmation deriving from the fact that a number of models agree on a result. Instead, I want to offer some suggestions here of why economists may want to engage in these practices.

The first case occurs when we are dealing with de-idealisation: Here economists want to show that a model which is strictly preferable for inferential purposes, as we have seen, yields the same result as an original model. Thereby, they may justify the use of the original model for explanatory purposes (unification) or for easier exposition (simplicity). However, as we just argued, any additional confirmation derives from the increased adequacy of the new model. There is no confirmation deriving from the fact that the two models agree, that a result is robust. The agreement is a symptom of the fact that often models have the purpose of explanation, and then a de-idealised model is only any good when it shares the result that represents the phenomenon to be explained.

The second case occurs when models are substantively different from each other, as we have said is often the case, and are meant to apply in different kinds of situations. Again, here we do not seem to get any confirmation out of the fact that models agree. How well we can make inferences from each of these models depends on how well they each fit the situations in which they are meant to apply.

So what is the special significance in having models that agree with each other, when they are each meant to apply in different circumstances? As we have seen, economists go to some lengths to engineer models that will furnish some result (the non-experimental nature of robustness analysis). Two reasons stand out why it could be important or significant to have substantively different models agree on some result.

Firstly, as our discussion of de-idealisation showed, in the case where a highly idealised baseline model can be de-idealised in a number of ways to fit different kinds of situations, we may get some explanatory unification from the baseline model. Here we would be dealing with models which are substantively different in a number of ways, but which share some features in the causal mechanism that brings about the result of interest. The baseline model would be serving the role of highlighting the features the different models have in common and could offer some explanatory unification. But for this to be the case, we need the models to all agree on some result that represents the explanandum, which is where robustness comes in.

This is exactly what seems to be going on in Hirshleifer and Teoh's collection of models of informational cascades. The different models they discuss all in a sense take Banerjee's or Bikhchandani et al.'s models as a baseline case, and present causal mechanisms that are substantively different but all share the idea that some form of imperfect information makes people make inferences from the observation of other agents' behaviour which causes herding on some choice or belief, frequently a bad choice or a false belief. If we were to explain a particular instance of herd behaviour, or if we wanted to make a prediction in a particular case where we think imperfect information of the type required may be present, then we would probably want to pick one specific model which best fits the situation we are interested in. This model may be based on Banerjee's baseline model, but will have been adjusted to fit the specific instance of herd behaviour, that is de-idealised in the right way, for the case we are interested in. So if we are interested in financial markets where firms can delay action, and partly what we want to explain is the rapid nature of an informational cascade, and its exact timing, then Chamley and Gale's (1994) model will serve best – which is a model that has replaced the assumption of a predetermined order of sequential decisions by the assumption that agents choose when to make a choice. The benefit from having a baseline model like Banerjee's, and treating all the models that share a "robust" result as a family, is that it makes us see similarities between cases, establish herd behaviour as a phenomenon that occurs in a variety of different situations, and lets us offer explanations that are more unified. In this case robustness just is about having a variety of different models that can potentially explain a phenomenon to pick and choose from, and our ability to offer explanations that are more unified. There is nothing about robustness that confirms particular inferences from models: We need to already be confident that each of the models applies to a particular situation.

The second reason for aiming at a variety of substantively different models that yield the same result is that this may teach us about robustness in the target, which brings us back to an important distinction made in chapter 3. If each of the substantively different models applies to a different kind of situation in the target system, that is, if we can make inferences from each of them about some aspects of the target, then showing that many different models yield the same result means showing that there is some kind of robustness in the target. By accumulating models of herd behaviour that are each meant to apply in slightly different kinds of situations, we could show that herd behaviour actually does occur in a variety of situations, that it is a common feature of the economic or social realm. Or, if we have a number of models that show that a particular informational structure leads to herd behaviour, we could show that this is a causal relationship which is robust in the sense that it occurs in a variety of different actual settings.

As we have seen in chapter 3, robustness in the target is a useful thing to know about for a number of reasons. Stable causes may be part of the nature of the social realm and just an important thing to know about when we want to understand it. They may also help us when we want to extrapolate from one context to another: Knowing that herd behaviour can occur in a variety of situations may make us more confident that it occurs in a new context. And showing that a phenomenon occurs in a variety of different situations may just be an important thing to show in order to demonstrate that the phenomenon is a worthwhile thing to study. After all, economists often spend a good amount of time when introducing a model on listing the variety of different kinds of situations their model may have something to say about. Banerjee does exactly that: He talks about fashions, restaurants and financial markets before introducing his actual model. What he wants to say is that herd behaviour is a wide-spread phenomenon and hence a fruitful topic for investigation. When in the wake of the presentation of his model, many more models of herd behaviour, that are adapted to fit different kinds of scenarios, are constructed, this may be understood as a way of showing that in fact, herd behaviour is a wide-spread phenomenon, not just in the model world, but also in the target system.

In this case, then, we make an inference from robustness in the model system – the fact that there are many substantively different models that yield the same result – to robustness in the target – that that result in fact occurs in a variety of different settings in the target. But this is very different from using robustness in the model system – which is a form of methodological robustness, where what is robust is the result of a scientific investigation – to improve one particular inference from the model to a target, or to confirm one single hypothesis about the target. Each of the models that yield the same result is thought of as telling us something about the target - each individually confirms a hypothesis about the target. Taking all these hypotheses together, we get a claim about some form of stability in the target, about the prevalence of a phenomenon such as herd behaviour. But here in fact we already need to be confident in each of the models in our target to allow for reliable inferences about the target. Only then can we conclude from methodological robustness (robustness in the model system) to robustness in the target. But robustness tells us nothing about the reliability of each of these inferences. Robustness is interesting because we think it translates into the target, but it is not confirmatory in the way the other two forms of methodological robustness envisage.

7.3. Conclusions

What I suggest economists are aiming to do when they derive the same result using alternative models is either to de-idealise and simply come up with a model which is preferable for inferential purposes, or to make a model fit different kinds of contexts, resulting in models that apply in different situations in the target. They are not using collections of models to improve one particular inference about the target, or for the models to collectively support one hypothesis about the target, as both robustness analysis as inferential robustness and as agreement of a variety of evidence would suggest.

When economists de-idealise a model, they do not only come up with a model that is preferable for inferential purposes, but they may also want to keep an old model and confirm its use for explanation: An original baseline model may both have the virtue of simplicity, and of offering unification of a number of diverse instances of phenomena like herd behaviour. Unification may also be a major motivation when economists create models that are substantively different from each other, but can all be related back to some baseline model like Banerjee's. In both these cases, the motivation for accumulating models that share a core result lie more in the purpose of explanation rather than confirmation.

Finally, when accumulating substantively different models, economists may be mostly interested in robustness in the target - in how prevalent a certain phenomenon is in the target, and the variety of circumstances under which a phenomenon such as herd behaviour occurs. In this case methodological robustness – the agreement of a variety of models – is used as evidence for robustness in the target, rather than as a way of confirming a particular hypothesis. In this case, there is nothing so special from an epistemic point of view in what is going on in robustness analysis. A variety of models is used to learn about a variety of situations in a target. The role of robustness analysis in providing simpler and more unified explanations on the other hand gives a more significant role to the fact of the agreement of models, and may be worth investigating more.

We have seen, then, that economists have different goals in mind than confirmation, and indeed goals which seem less unattainable. While these ideas need developing, and not all of the goals I cited will turn out defensible, at least our observations point to a more fruitful way to approach the practice of robustness analysis - or derivational robustness to use Woodward's term.

8. In Conclusion

This thesis has been concerned with the practice of robustness analysis in theoretical economics and the epistemic hopes some philosophers have attached to it. Driven by unsettled questions about how we can learn from theoretical models in economics, some have thought that robustness analysis is a way, if not the most important way, to acquire confidence in inferences from highly idealised models. What we have shown is that these hopes are unwarranted. We have distinguished two arguments that have been made to the effect that robustness analysis can confirm hypotheses about a target, and rejected both with the aid of a case study on models of herd behaviour.

In the first argument, we conceive of the models compared in robustness analysis as independent pieces of evidence for a hypothesis, whose agreement would be too much of a coincidence if the hypothesis whose truth they all indicated wasn't true. The main requirement for this kind of argument form robustness is that the different sources of evidence are independent. But we have shown that independence usually fails: The main problem with this argument is that the models that are in fact compared in robustness analysis either share many assumptions, which introduces shared biases as another plausible explanation for their agreement, or they have been selected according to, or designed in order to deliver the robust result – in which case selection is the most plausible explanation for agreement.

In the second argument, which is a version of Woodward's 'inferential robustness', we use robustness analysis to make judgements about which assumptions do and do not drive a result. Robustness analysis can tell us that a particularly worrisome assumption is irrelevant for the derivation of a result. This is then taken to warrant increased confidence in whatever inference we wanted to make from the model. The main requirement for such an argument to work is the exhaustiveness of alternatives tried. Again, we show that this requirement is typically not met: The problem with this argument is that we can usually not conclude from robustness analysis that particular assumptions are unimportant for the derivation of a result – this would require trying out a larger variety of alternatives than we usually can, and doing so would in most cases result in the failure of robustness. Further, even if we did find this out, this does not generally license increased confidence in inferences from the model. This is also due to the fact that whether an assumption is relevant, or whether its being 'unrealistic' is cause for worry about inferences from the model can depend on the rest of the model.

What the main case study of this thesis also shows, however, is that economists are not in fact after confirmation when they accumulate alternative models that share a core result with an original model. So while the philosophical accounts of the value of robustness analysis discussed in this thesis fail, this does not need to mean that the practice of robustness analysis is pointless. The last chapter explored some alternative interpretations of what economists may in fact be aiming for when they conduct robustness analysis.

Two open questions motivated us in the introduction: We were concerned about how to learn from idealised economic models, and we were wondering about the use of the practice of robustness analysis. Now we have found that one does not provide a solution to the other: Robustness analysis does not solve the problem of unrealistic assumptions. Yet, as chapter 7 showed, this does not need to mean that the practice of robustness analysis is pointless. In fact there are many interesting potential uses of robustness to be explored, if we only look closer at modelling practice and the wider literature on modelling.

Woodward (2006) stressed that the robustness of results in theoretical modelling, what he calls derivational robustness, was different from other kinds of methodological robustness. Most of this thesis dealt with arguments that claimed that robustness in modelling could be confirmatory just in the same way as methodological robustness when it comes to measurement, or to inferences from sets of data. Now that we have seen the failures of these arguments, and after having had a closer look at what economists seem to be doing when they conduct robustness analysis, it seems that Woodward was right after all[7].

Talk of robustness in theoretical modelling is not like appeals to robustness in other areas of science, for instance when it comes to measurement. In this thesis, we could only touch on what exactly the purpose of robustness analysis could in fact be. Instead of nourishing the hope that somehow we could get confirmation out of robustness analysis, and solve the problem of learning from models with unrealistic assumptions, it may be more fruitful to investigate the role that the derivation of a result via multiple models plays in explanatory unification, and in establishing the wide-spread nature of a phenomenon.

[7] Interestingly, he even noted that robustness analysis in theoretical modelling often concerns the variation of what he calls values of parameters, but which we can understand as substantive assumptions – this being one of the main reasons we identified in the last chapter why much of robustness analysis cannot be understood as aiming for confirmation:

> *"In the latter case, in which the concern is with derivational robustness, an assumption is adopted about the value of the parameter and this is used, in conjunction with other theoretical assumptions, to derive some range of observed phenomena. Investigations are then made whether, given other values of the parameter, but the same theoretical assumptions, the same conclusions can be derived." (p.233)*

Appendix:

To convince the reader that the results from our case study are not just specific to the case we introduced, let me briefly look at the case study provided by Kuorikoski et al. (2010), which if anything would be expected to make the strongest case in favour of their argument. However, as we will see, their case illustrates many of the problems our case study on herd behavior illustrated.

The core model Kuorikoski et al. use is a model in geographical economics, namely Krugman's (1991) "Increasing Returns and Economic Geography". This model explores the conditions under which there is spatial agglomeration of industries in a country, and an industrial core and an agricultural periphery develop. The following are the key assumptions of the model:

Krugman 1991:

- There are only 2 regions, and the regions themselves do not have geographical extension.
- Production:
    - There are only 2 kinds of production (agriculture and manufacturing).
    - Agricultural production:
        - Constant returns
        - Factor immobility
    - Manufacturing:
        - Increasing returns (fixed cost plus constant marginal cost)
        - Factor mobility
- Industrial Organisation:
    - Monopolistic competition:
        - Large number of firms
        - Free entry
        - Product differentiation
        - Each firm assumes to have a negligible effect on the market, so there is no strategic interaction.
        - Profit maximisation
- Demand side:
    - Utility maximisation
    - All consumers have the same utility.
    - Utility depends only on the consumption of agricultural and manufacturing goods.
    - CES (constant elasticity of substitution) utility function with constant shares for agriculture and manufacturing; within manufacturing there is a preference for variety
- Transportation:

    - Costless for agricultural products
    - "Iceberg" costs for manufacturing: a constant share of the products gets lost on the way from one region to the other
- The model is solved for equilibrium

The main results from Krugman's model are that manufacturing concentrates in one region if

- transportation costs are low,
- the constant share of manufacturing consumed is high, and
- there are large economies of scale.

Kuorikoski et al. give one main example of a model that demonstrates the robustness of the core results of Krugman's model, namely Ottaviano et al. (2002). This model shares all the assumptions of Krugman's original model, making the following alterations:

1) Utility: Utility is still identical across individuals, but it is quasi-linear with a quadratic and symmetric sub-utility for variety of manufacturing goods.
2) Transportation costs: There is no melting away, but a fixed unit cost.
3) They use a different equilibrium concept, which allows them to solve the model analytically.

With these assumptions, they claim to be reproducing the same results as Krugman (1991): Agglomeration occurs with low transportation costs, a high share of manufacturing and larger economies of scale.

For several reasons, this case study is not a good example for the confirmatory value of robustness analysis as inferential robustness, and in fact just reproduces much of what we have said with our case study on herd behaviour:

Firstly, several of Ottaviano et al.'s changes could be seen as de-idealisations rather than as the variation of tractability assumptions. At least they themselves seem to think that they are replacing unrealistic assumptions with more realistic ones:

> *"Taken together, these [Krugman´s] assumptions yield a demand system in which the own-price elasticities of demands are constant, identical to the elasticities of substitutions and equal to each other across all differentiated products. This entails equilibrium prices that are independent of the spatial distribution of firms and consumers. Though convenient from an analytical point of view, such a result conflicts with research in spatial competition which shows that demand elasticity varies with distance while prices change with the level of demand and the intensity of competition. Moreover, the iceberg assumption also implies that any increase in the price of the transported good is accompanied by a proportional increase in its trade cost, which is unrealistic. Third, the stability analysis used to select spatial equilibria rests on myopic adjustment processes in which the location of mobile factors is driven by differences in current returns."* (p.410)

Secondly, many of the assumptions in Krugman's original model seem to be unrealistic, so that it is not clear what trying out one alternative specification that changes only three of them should tell us. In particular, the assumptions that we are dealing with just two regions of no spatial extension, that there is free entry to the market, and that there is no cost to the transportation of agricultural goods are not changed. This means that all the problems we talked about in the context of Banerjee's model apply: Robustness analysis is partial and non-exhaustive, even to a seemingly stronger extent than the robustness analysis we looked at in the case of herd behaviour.

Thirdly, Ottaviano et al. themselves do not seem to have the goal of strengthening our belief in some hypothesis about the target. Their expressed goals are quite different. They claim to develop an alternative model because they see benefit in having a model that is analytically solvable. Again, as with de-idealisation, they simply think that their model is better than Krugman's, not that their model and Krugman's in combination allow for deriving a more reliable robust theorem. Only here, in contrast to de-idealisation, the advantages are not meant to be necessarily epistemic: An analytically solvable model allows for more clear-cut comparative statics results, as well as welfare analysis, which Krugman's model couldn't deliver. Not only does this mean that Ottaviano et al. did not have the goals Kuorikoski et al. would like to ascribe to them, but these alternative goals also mean that Ottaviano et al. may have specifically designed the model to reproduce Krugman's results, in order to reap these benefits. As we have seen, this element of design can be a problem for accounts that see confirmatory value in robustness analysis.

References:

ALDRICH, J., 2006. “When are Inferences too Fragile to be Believed?”, *Journal of Economic Methodology,* 13 (2), pp. 161-177.

AKERLOF, G., 1970. “The Market for 'Lemons': Quality Uncertainty and the Market Mechanism” , *The Quarterly Journal of Economics,* 84 (3), pp. 488–500.

ALEXANDROVA, A., 2008. “Making Models Count”, *Philosophy of Science*, 75, pp. 383-404.

BAKER, A., 2011. "Simplicity", *The Stanford Encyclopedia of Philosophy (Summer 2011 Edition)*, in ZALTA, E. (ed.), URL = http://plato.stanford.edu/archives/sum2011/entries/simplicity/.

BANERJEE, A., 1992. “A Simple Model of Herd Behavior”, *The Quarterly Journal of Economics,* 107 (3), pp.797-817.

BARNETT, T., PIERCE, D., HIDALGO, H., BONFILS, C., SANTER, B., DAS, T., BALA, G., WOOD, A., NOZAWA, T., MIRIN, A., CAVAN, D., and DETTINGER, M., 2008. “Human-Induced Changes in the Hydrology of the Western United States”, *Science,* 319 (5866), pp.1080-1083.

BARRO, R., and JIN, T., 2011. “On the Size Distribution of Economic Disasters”, *Econometrica,* 79 (5), pp. 1567-1589.

BEGHEIN, C., and TRAMPERT, J., 2003. “Robust Normal Mode Constraints on Inner-Core Anisotropy from Model Space Search”, in *Science,* 299 (5606), pp. 552-555.

BERTRAND, J., 1883. "Book review of theorie mathematique de la richesse sociale and of recherches sur les principles mathematiques de la theorie des richesses", *Journal de Savants* (67), pp. 499–508.

BIKHCHANDANI, S., HIRSHLEIFER, D., and WELCH, I., 1992. “A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades”, *Journal of Political Economy*, 100, pp. 992-1026.

BOUMANS, M., 2005. *How Economists Model the World into Numbers*, London and New York: Routledge.

BOVENS, L., and HARTMANN, S., 2003. *Bayesian Epistemology.*

BURGUET, R. and VIVES, X., 2000. “Social Learning and Costly Information Acquisition”, *Journal of Economic Theory*, 15, pp. 185-205.

CAO, H., and HIRSHLEIFER, D., 2000. “Conversation, Learning and Informational Cascades”, *Ohio State University Fisher College of Business Working Paper*.

CARTWRIGHT, N., 1983. *How the Laws of Physics Lie*.

CARTWRIGHT, N., 1991. "Replicability, Reproducibility, and Robustness: Comments on Harry Collins", *History of Political Economy*, 23 (1), pp. 143-155.

CARTWRIGHT, N., 1999. “Capacities” in DAVIS, J., HANDS, W., and MÄKI, U., (eds.) *The Handbook of Economic Methodology*.

CARTWRIGHT, N., 2006. "The Vanity of Rigour in Economics: Theoretical Models and Galilean Experiments", in FONTAINE, P., and LEONARD, R. (eds.), *The 'experiment' in the history of economics*.

CARTWRIGHT, N., 2009. “Causality, Invariance and Policy”, in KINCAID, H., and ROSS, D. (eds.), *The Oxford Handbook of Philosophy of Economics*.

CHAMLEY, C., and GALE, D., 1994. “Information Revelation and Strategic Delay in Irreversible Decisions”, *Econometrica*, 62, pp. 1065-1085.

CHARI, V. and KEHOE, P., 2000. “Financial Crises as Herds”, *Working paper, Federal Reserve Bank of Minneapolis* # 600.

COLLINS, H., 1991. “Replicability, reproducibility, and robustness”, *History of Political Economy*, 23 (1), pp. 23–42.

ELLISON, G., and FUDENBERG, D., 1993. “Rules of Thumb for Social Learning”, *Journal of Political Economy*, 101, pp. 612-643.

FRIEDMAN, M., 1953. “The Methodology of Positive Economics”, in his *Essays in Positive Economics.*
GUALA, F., and SALANTI, A., 2002. “On the Robustness of Economic Models”, *Working Papers 0208, University of Bergamo, Department of Economics.*
GUALA, F., 2005. *The Methodology of Experimental Economics.*
HARTWELL, L., 2004. “Robust Interactions”, in *Science*, 303 (5659), pp.774-775.
HEMPEL, C., 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: Free Press.
HIRSHLEIFER, D., and TEOH, S., 2003. “Herd Behavior and Cascading in Capital Markets: A Review and Synthesis”, *European Financial Management*, 9(1), pp. 25-66.
HOOVER, K., and PEREZ, S., 2004. “Truth and robustness in cross-country growth regressions”, *Oxford Bulletin of Economics and Statistics*, 66, pp. 765–98.
HOOVER, K., 2006. “Fragility and Robustness in Econometrics: Introduction to the Symposium”, *Journal of Economic Methodology,* 13 (2), pp. 159-160.
ISMAEL, J., 2009. "Quantum Mechanics", ZALTA, E. (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <http://plato.stanford.edu/archives/fall2009/entries/qm/>.
KITCHER, P., 1989. “Explanatory Unification and the Causal Structure of the World”, in KITCHER, P., and SALMON, W., *Scientific Explanation*.
KRUGMAN, P., 1991. “Increasing Returns and Economic Geography”, *The Journal of Political Economy,* 99 (3), pp. 483-499.
KUORIKOSKI, J., LEHTINEN, A., and MARCHIONNI, C., 2010. “Economic Modelling as Robustness Analysis”, *British Journal for the Philosophy of Science*, 61 (3), pp. 541-567.
LEAMER, E., 1983. “Let’s take the con out of econometrics”, *American Economic Review,* 73, pp. 31-44.
LEAMER, E., 1985. “Sensitivity Analysis Would Help”, *American Economic Review,* 75, pp. 308-313.
LEVINS, R.,1966. “The strategy of model building in population biology” in SOBER, E., (ed.), *Conceptual Issues in Evolutionary Biology.*
LEVINS, R., 1968. *Evolution in Changing Environments: Some Theoretical Explorations.*
LOMBROZO, T. and RUTSTEIN, J., 2004. “Simplicity in Explanation”, *Proceedings of the 26th annual conference of the Cognitive Science Society*, pp. 837-842.
MILGROM, P., and ROBERTS, J., 1981. “Predation, Reputation, and Entry Deterrence”, *Journal of Economic Theory*, 27 (2), pp. 280-312.
MILGROM, P., and ROBERTS, J., 1982. “Limit Pricing and Entry under Incomplete Information: An Equilibrium Analysis”, *Econometrica*, 50 (2), pp. 443-459.
MORGAN M., and MORRISON, M., 1999. *Models as Mediators*. Cambridge: Cambridge University Press.
ODENBAUGH, J., and ALEXANDROVA, A., 2011. “Buyer Beware: Robustness Analyses in Economics and Biology”, *Biology & Philosophy*, 26 (5), pp. 757-771.
ORZACK, S., and SOBER, E., 1993. "A Critical Assessment of Levins's the Strategy of Model Building in Population Biology (1966)", *Quarterly Review of Biology*, 68 (4), pp. 533-546.
OTTAVIANO, G., TABUCHI, T., and THISSE, J., 2002. “Agglomeration and Trade Revisited”, *International Economic Review,* 43 (2), pp. 409-436.
PARKER, W., 2011. “When Climate Models Agree: The Significance of Robust Model Predictions”, *Philosophy of Science,* 78 (4), pp. 579-600.
REISS, J., 2007. *Error in Economics: Towards a More Evidence-Based Methodology.*
SELTEN, R., 1978. "The chain store paradox". *Theory and Decision* 9 (2), pp. 127–159.
SCHELLING, T., 1969. "Models of segregation", *American Economic Review*, 59(2), pp. 488-493.
SIRCAR, R., and LEDVINA, A., forthcoming. “Static and Dynamic Oligopoly Games under Asymmetric Costs”

SPENCE, A., 1977. “Entry, Capacity, Investment in Oligopolistic Pricing”, *The Bell Journal of Economics*, 8 (2), pp.534-544.
STEGENGA, J., 2009. “Robustness, Discordance, and Relevance”, *Philosophy of Science,* 76 (5), pp. 650-661.
SUGDEN, R., 2000. “Credible Worlds: The Status of Theoretical Models in Economics”, *Journal of Economic Methodology*, 7 (1), pp. 1-31.
SUGDEN, R., 2009. “Credible Worlds, Capacities, and Mechanisms”, *Erkenntnis*, 70 (1), pp. 3-27.
VARIAN, H., 2002. *Intermediate Microeconomics: A modern Approach,* 6th edition.
VIVES, X., 1993. “How Fast Do Rational Agents Learn?”, *Review of Economic Studies*, 60, pp. 329-347.
WAYNE, A., 1995., “Bayesianism and Diverse Evidence”, *Philosophy of Science,* 62 (1), pp. 111-121.
WEBB, E., CAMPBELL, D., SCHWARTZ, R., and SECHREST, L., 1966. *Unobstrusive Measures: Non-reactive Research in the Social Sciences.*
WEISBERG, M., 2006a. "Forty Years of 'the Strategy': Levins on Model Building and Idealization", *Biology and Philosophy*, 21, pp. 623-645.
WEISBERG, M., 2006b. "Robustness Analysis", *Philosophy of Science*, 73 (5), pp. 730-742.
WEISBERG, M., and REISMAN, K., 2008. “The Robust Volterra Principle“, *Philosophy of Science*, 75, pp. 106-131.
WIMSATT, W., 1981. "Robustness, Reliability and Overdetermination", in BREWER, M., and COLLINS, B. (eds.), *Scientific inquiry and the social sciences.*
WIMSATT, W., 1987. "False Models as Means to Truer Theories", in NITECKI, M., and HOFFMAN, A. (eds.), *Neutral Models in Biology.*
WIMSATT, W., 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.*
WOODWARD, J., 2003. *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.
WOODWARD, J., 2006. "Some Varieties of Robustness", *Journal of Economic Methodology*, 13 (2), pp. 219-240.