Johanna Thoma

On the hidden thought experiments of economic theory

On the Hidden Thought Experiments of Economic Theory

Johanna Thoma[1]

Abstract: Most papers in theoretical economics contain thought experiments. They take the form of more informal bits of reasoning that precede the presentation of the formal, mathematical models these papers are known for. These thought experiments differ from the formal models in various ways. In particular, they do not invoke the same idealized assumptions about the rationality, knowledge and preferences of agents. The presence of thought experiments in papers that present formal models, and the fact that they differ from the formal models in this way are often ignored in debates on what, if anything, we can learn from formal models in theoretical economics. I show that paying due attention to thought experiments in theoretical economics has serious implications for this debate. Differences between thought experiments and formal models are especially problematic for Sugden's 'credible worlds' account.

1. Introduction

Economics is known as a science of mathematical models. A typical model in theoretical economics is a deductive argument, which derives a conclusion from a set of mathematically expressed assumptions. It features agents who face some choice problem, and who are assumed to be perfectly rational, in the sense that they abide by the tenet of rationality axioms of economic utility theory concerning both their choice of action and their belief formation. It is also likely that the model will assume that the agents featured in the model are greedy, in the

[1] I am very grateful to Jim Brown, Philippe Verreault-Julien and François Claveau for extensive comments on earlier drafts of this paper. Margaret Schabas and an audience at RT/ENPOSS at the University of Washington also provided helpful feedback.

sense that they always prefer more of any given good, and selfish, in the sense that other people's wellbeing does not matter to them directly. Indeed, a great number of models in economics assume that all that the agents featuring in them care about is money. And lastly, it will make some idealized assumptions about what the agents know about their environment and other agents. This combination of characteristics, or a subset of them, is what critics of economics seem to have in mind when they speak of 'Homo oeconomicus'.

Most people violate the rationality axioms of economic utility theory. Furthermore, real agents are not always greedy and they care about more than money, in particular about the wellbeing of others. Thus, the types of agents featured in economic models are not 'realistic'. Still, ideally, we would like these models to apply to a real world 'target', in the sense that they either explain some real world phenomenon, or make some prediction about real life markets or human interactions. Now if there is a mismatch between the kinds of agents we find in economic models and the agents who operate in the target, the model is 'unrealistic', and it is at first sight unclear how the model can tell us anything about the target.[2]

The question of what the epistemic status of these unrealistic models is - whether and how we can learn from them about real world targets - is extremely controversial in the philosophy of economics. What is often overlooked in this debate is the fact that in papers in theoretical economics, formal models of the type I just described are usually preceded by a bit of more informal reasoning that is more aptly described as a *thought experiment*. While being less formally rigorous, these thought experiments are more 'credible' than the formal models in a way that I will explain below. As I want to show in this paper, paying closer attention to this fact causes trouble for the most prominent accounts of the status of 'unrealistic' models.

To start, the fact that these thought experiments are presented in the exposition of a formal model tells us something about the aspirations economists have when they construct models: The thought experiment concerns a choice situation that could be real, and employs the

[2] Economic models tend to be unrealistic in other ways, too. In particular, the kinds of choice situations and markets that feature in them are often extremely simple compared to real life markets.

reasoning patterns real agents may use. It thus seems like economists think they are in the business of directly explaining real world phenomena, contrary to some accounts of the purpose of models in theoretical economics.

Amongst those who have recognized the importance of the informal stories that accompany models,[3] it is common to assume that these stories are just an integral part of the model. But a closer look at these thought experiments brings to light important differences between the thought experiment and the formal model, and this paper explores some of the implications of this observation. While nothing I will say speaks determinately against defining models as consisting of a formal model together with an informal story, what I hope to show is that this way of thinking about models may cover up some important problems.

The differences between model and thought experiment cast doubt on one recent and influential account of how we learn from economic models that does take seriously the ambitions revealed in the thought experiments that precede formal models. Robert Sugden's 'credible worlds' account (see Sugden 2000, 2009) claims that economists construct alternative worlds that act as analogies to various real world phenomena. We are licensed to make inferences from these analogous worlds in large part because we take them to be 'credible'.[4] Sugden thinks that the thought experimental stories that economists tell around their models establish this credibility. Since Sugden shares the common assumption that thought experiment and model are just variants of the same thing, the thought experiment's credibility transfers onto the formal model.

What I want to argue below is that this assumption is not warranted. The thought experiment thus does not establish the credibility of the model, and it is not clear what else could. Instead, I will suggest that a more plausible role for the thought experiment is twofold: It

[3] E.g. Gibbard and Varian 1978, Hartmann 1999, Morgan 2001, Grüne-Yanoff and Schweinzer 2008, Alexandrova 2009.

[4] In his 2013, Sugden stresses that the key concept in his account is not credibility but similarity, and that credibility is neither necessary nor sufficient for a model to apply to a real world phenomenon. I take it to be fairly uncontroversial that some kind of similarity needs to hold between model and world. An account of how we can learn from models should ideally say more than that, and it is Sugden's appeal to credibility that allows him to do so. If his appeal to credibility is problematic, as I want to argue below, then this at the very least calls into question the core, and most distinctive part of his account.

acts as independent evidence for the hypothesis the model wants to establish, and it acts as an example of a more realistic case that the formal model does apply to, making it more convincing that it applies to real world phenomena. But exactly how the formal model applies to the thought experimental story or any real world phenomenon is still an open question.

I will proceed as follows: Section 2 will introduce a paradigmatic paper from theoretical economics that does consist of a thought experiment - formal model pair. Section 3 will compare the two and also explain why the former counts as a thought experiment while the latter does not. Section 4 draws out the implications for the debate on the epistemic status of ‘unrealistic’ models in economics, and calls into question Sugden’s account in particular. Section 5 concludes with a conjecture about the role of thought experiments as they are presented in the exposition of formal economic models.

2. A Thought Experiment about Restaurant Choice

The example I want to focus on here, Banerjee (1992), is an influential paper on a phenomenon sometimes referred to as ‘herd behavior’. It occurs whenever a group of people all behave in the same way even though they have received diverse information concerning what the best course of action is. Herd behavior is said to occur in a wide range of contexts, from financial markets to fashion and choice of research topic. An intuitive explanation of this phenomenon might be that people simply have a brute preference for doing what everybody else is doing. But Banerjee wants to show that it can occur even if such a preference is absent. In his paper, he presents an explanation of the phenomenon of herd behavior that attributes it to a kind of imperfect information: agents only have private signals about some object of choice, but try and infer something about other people’s information by observing their behavior. This seems intuitive especially in the case of financial markets: Everybody is to some extent in the dark – but if somebody else makes a choice, we may think they know something that we do not. As a result, we act like them, and become less responsive to our own information. In the end, collectively we do not use our information efficiently, and what we do depends on the information the first few agents to act happened to have.

Banerjee presents a formal model that illustrates this process, sometimes referred to as an *informational cascade*, which takes up most of the paper. But before turning to this model, he presents the following story about restaurant choice:

> Most of us have been in a situation where we have to choose between two restaurants that are both more or less unknown to us. Consider now a situation where there is a population of 100 people who are all facing such a choice.
>
> There are two restaurants A and B that are next to each other, and it is known that the prior probabilities are 51 percent for restaurant A being the better and 49 percent for restaurant B being better. People arrive at the restaurants in sequence, observe the choices made by the people before them, and decide on one or the other of the restaurants. Apart from knowing the prior probabilities, each of these people also got a signal which says either that A is better or that B is better (of course the signal could be wrong). It is also assumed that each person's signal is of the same quality.
>
> Suppose that of the 100 people, 99 have received signals that B is better but the one person whose signal favors A gets to choose first. Clearly, the first person will go to A. The second person will now know that the first person had a signal that favored A, while her own signal favors B. Since the signals are of equal quality, they effectively cancel out, and the rational choice is to go by the prior probabilities and go to A.
>
> The second person thus chooses A regardless of her signal. Her choice therefore provides no new information to the next person in line: the third person's situation is thus exactly the same as that of the second person, and she should make the same choice and so on. Everyone ends up at restaurant A even if, given the aggregate information, it is practically certain that B is better. (pp. 798-799)

This story takes the reader through a choice situation that is similar to one that she might have faced before. Several features of the situation are perhaps a little artificial: Everybody faces the same choice of two restaurants, and somehow all agree that one of them is likely to be marginally better than the other. Moreover they all receive signals of the same quality that are nevertheless fallible. Here the reader needs to fill in the details for herself (Maybe they all read reviews from newspapers or websites of the same quality, but one of the reviews got it wrong). But even if the choice situation is somewhat artificial, Banerjee takes us through reasoning that we find plausible and familiar. Nothing is said about what kinds of agent these 100 agents are, so the presumption is that they are typical agents like you and I. The way the story is told makes it natural for us to imagine that we are in the choice situation the agents in the story face, and we find the behavior that Banerjee describes plausible. Only casual reference is made to rationality.

The formal model Banerjee presents in the following is less specific about the situations in which agents choose. For instance, no mention of restaurants or any other specific object of choice is made. Instead of facing a binary choice, the agents face a continuum of options. Only one option offers a return, which is the same for every agent. The agents do not receive a signal for certain, but only with a probability α, which tells the truth about which option offers the return with probability β, and gives a random, uniformly distributed signal otherwise. Agents choose in sequence and can observe all the choices made previously.

At the same time, the model makes a lot of assumptions about agents. This makes it possible to import the formal apparatus of economic utility theory, and game theory. There are N agents with identical, risk neutral von Neumann-Morgenstern utility functions (which value getting the return more highly than getting no return). They integrate new information in accordance with Bayes' Rule. Both the fact of their rationality and the setup of the situation are common knowledge amongst the agents. This means that each agent knows that all others are rational, and that all others know that she is rational, and that all others know that she knows that they are rational, and so on, ad infinitum. The model is then solved for a Bayesian Nash equilibrium.[5]

Banerjee uses the formal model to show the following: First, after the first few choices, agents will start to 'herd' on one choice. Second, what option 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, there is a high probability that

[5] For an explanation of all these formal concepts, see Mas-Colell, Whinston and Green (1995)

nobody chooses the ‘right’ option, even though many agents received the right signal. It is these results for which the paper is known amongst economists.

Banerjee’s model is no exception in combining an informal story about a real life market or interaction with a formal model in this way. Ever since Akerlof’s (1970) famous “Market for Lemons” paper, which also begins with an intuitive, informal story, most successful papers in theoretical economics have done.[6] The authors of these papers are usually not very explicit about what the stories are doing in their papers and how they relate to the model. The stories tend to be introduced as illustrations of the kind of thing that the model is taken to show. Note, however, that the story that I quoted above does not make any explicit assumptions about what kinds of agents we are dealing with. There is neither talk of Bayesian rationality, nor of common knowledge, nor of all the agents having identical preferences. The story is not explicitly a story about ‘Homo oeconomicus’. This, too, is a typical feature of the stories that economists tell before they introduce a formal model. I think these differences between the story and the formal model are significant enough that the first deserves to be called a ‘thought experiment’ while the latter does not, as I will explain in the next section. And, as I will then argue, the differences are also large enough to call into question an influential account of how we learn from formal economic models.

3. Models and Thought Experiments

It is sometimes said that models are a lot like experiments, or indeed that they are types of experiment. Cartwright (1999) claims that economic models are like experiments in

[6] Before then, this practice does not seem to have been as common. Akerlof received multiple rejections from journals before his paper was published, on the basis that nobody is interested in markets for used cars. Akerlof (2003) remembers: “I received my first rejection letter from *The American Economic Review.* The editor explained that the *Review* did not publish papers on subjects of such triviality. In a case, perhaps, of life reproducing art, no referee reports were included.” Akerlof later received the Nobel Memorial Prize in Economics primarily for this paper. The point is that Akerlof did not mean for the paper to be *about* the market for used cars, at least not primarily. In the same way, Banerjee’s paper is not *about* restaurant choice. But such ‘trivial’ examples have come to play an important role in economic theorizing.

physics. First, both involve starting from an idealized set-up. In the case of the economic model, the kinds of unrealistic assumptions I described above constitute idealizations. In physics experiments, the idealization consists in creating ideal conditions in which to conduct the experiment, such as creating a near vacuum, or using planes that are as frictionless as possible. Further, both experiments and models then involve manipulation and the observation of the effects of the manipulation. Because of such similarities, Guala (2002) claims that models and experiments are tokens of the same kind. Mäki (2005) goes so far as to say that models are types of experiments. The only general difference between models and more conventional experiments is that the former are theoretical representations whereas the latter are material representations.

It is common for commentators who liken models to experiments to claim offhand that models are *thought* experiments. If models are experiments, but no material manipulation is involved, it seems natural to describe them as thought experiments to distinguish them. Still, considering the differences between Banerjee's story, and Banerjee's formal model draws attention to a feature that paradigmatic thought experiments have and that formal economic models do not.

Both the formal model and the informal story we looked at share some standard features of (thought) experiments. They start by describing a setup that is to some extent idealized or hypothetical. The story asks us to consider a hypothetical choice situation, and the formal model makes some idealized assumptions. Moreover, the initial setups in the two cases bear some resemblance. There is then a manipulation: In both cases, we are asked to consider what happens if we let a number of agents make choices in sequence. We then observe a thought experimental/model phenomenon: Herd behavior occurs. And note that the phenomenon is the same in both cases. Lastly, in both cases, if our aim is to explain real life cases of herd behavior, there is still an inductive leap to be made from the thought experimental/model phenomenon to the real world phenomenon. These are the similarities between the story and the formal model, and also features that are often ascribed to thought experiments.[7] Thought experiments start with some hypothetical conditions, involve manipulation, and produce a thought experimental phenomenon.

Still, there are important differences between the model and the story. First, note that in the model, the result is derived by calculation, using pen and paper or a computer. In the story, on the other hand, the phenomenon is established purely in the imagination. That a thought experimental phenomenon needs to be established by the imagination is something that Brown (1991, chapters 1&2) highlights for the case of thought experiments in natural science. He claims that the manipulations in thought experiments need to be mental manipulations. In Galileo's falling bodies thought experiment, which Brown treats as a paradigmatic case of a thought experiment, no calculations with pen and paper are necessary to see that the two bodies must be falling at the same speed. Imagining the situation is enough. If something similar is required of thought experiments in the social sciences, then only Banerjee's informal story seems to meet that standard.

But now maybe an economic modeler could say in response that she can simply carry out the relevant mathematical operations in her head, and that thus deriving the model result is a mental operation, too. She might even say that many years of experience in economic modeling have enabled her to see what the result will be simply from entertaining the assumptions of the model, and without even carrying out the operations explicitly in her head. Relatedly, some philosophers have suggested that modeling can be seen as 'extended cognition', rejecting the idea that there is a difference in kind between manipulations carried out using a material tool and manipulations carried out in the head.[8]

In response, we could say that there is more to imagination than performing a mental operation. A proper thought experiment may also need to be visualizable - which is also a feature Brown highlights. In Banerjee's story, we can imagine ourselves in the relevant choice situation, and visualize what it would be like to stand in front of the two

[7] See, for instance, Sorensen (1992).

[8] See, for instance, Giere (2002), Kuorikoski and Lehtinen (2009) and Toon (forthcoming).

restaurants, one of which is gradually filling with people. The formal model does not ask us to visualize the situation like this, while it is crucial for the story to be convincing.

But perhaps the model could be recast in a way that engages the imagination. The modeler would then presumably ask us to imagine that we were perfectly rational agents with common knowledge of the situation and others' rationality. Or, if we do not want to cast things in first personal terms, she would engage our intuitions about what these kinds of agents would do.

Here, I think, we reach the limits of our imagination. It is not clear if we have any intuitive access to what agents who are perfectly rational in this sense would do, or what it would be like to be one. In particular, common knowledge is something that is not humanly attainable: As we saw above, it is a shorthand for an infinite number of mental states. And it is at least not realistic that humans can have an infinite number of mental states. But even the other rationality assumptions seem out of reach considering that all the relevant rationality axioms are frequently violated (for an overview, see Kahnemann and Thaler 2006). Moreover, the preferences of these agents are impoverished compared to the agents we are familiar with, which makes it hard to imagine what they would do or what it would be like to be them.

It might be said that trained economists, having constructed many models of this type, do have fairly reliable third personal intuitions about what perfectly rational agents would do. But I think that these are not the kind of intuitions that classical thought experiments would appeal to. The kind of experience trained economists have is experience in applying economic utility theory. The kinds of agents that feature in economic models are theoretical constructions. And thus, formal models are mostly applications of this theory. But classical thought experiments, like Galileo's falling bodies thought experiment, are not merely applications of a theory, or do not merely involve a 'theory-based calculation' as Brown (1991, p.1) puts it. They are meant to engage intuitions that are fairly universal, and that we hold independently from some theory. The thought experimental phenomenon should be plausible or 'intuitive' for anybody who understands the setup of the thought experiment. But this is not so in the case of formal economic models: they are applications of a theory of rational choice, and moreover one that departs in important ways from 'folk psychology'.

The story that precedes Banerjee's formal model does seem to be plausible and intuitive independently of a theory of rational choice, and thus, I think, deserves to be called a thought experiment. We can imagine people reasoning in the way Banerjee describes here even if they are not perfectly rational or have common knowledge of the situation, and even if they have slightly different preferences from one another. It is plausible that real agents act in this way, and not just perfectly rational ones, and thus we can establish the thought experimental phenomenon simply by engaging our intuitions about how people behave.

Sorensen (1992) argues that we have a special strength when it comes to answering deliberative questions, and that many successful thought experiments play to this strength:

> Deliberative fields (ethics, law, economics) have an advantage because their hypotheticals already have a decision making format. The exercises in physics textbooks manifest appreciation of the point. They often put a practical spin on a hypothetical by making the problem one of escape, optimization, or strategy. In any case, nearly any thought experiment is improved by being pitched toward our practical side. (p. 39)

By featuring artificial agents that have mental capacities that are beyond those of normal people, as well as concerns that are narrower than those of normal people, formal economic models do not engage the reader's strength in answering deliberative questions - while the story about restaurant choice does.

There is one way, however, in which Banerjee's story does differ from classical thought experiments in physics, and that is that the thought experimental phenomenon does not have the same claim to necessity that, for instance, Galileo's falling bodies thought experiment has. Even though we think it is plausible that agents may reason in the way described, they need not. Perhaps there are situations where everybody's preferences are so idiosyncratic that nobody takes other people's choices as evidence for what is better for themselves. In that case the thought experimental phenomenon would not occur. The thought experimental phenomenon is plausible because we find the reasoning of the agents typical, not because we think they have to act in the way they do.

Nevertheless, Banerjee's story seems similar enough to paradigmatic cases of thought experiments to be called one. And I have argued that formal economic models are not.[9] While I think this is interesting in itself, the issue is not merely one of labeling. The presence of thought experiments in papers in theoretical economics, as well as the differences between these thought experiments and the formal models they accompany that we discussed here have some important implications for old methodological questions. In particular, they have implications for the debate on what, if anything, we can learn from formal models in theoretical economics. The next section describes two major lessons: First, the thought experiments reveal that theoretical economists have higher ambitions than most accounts of what economic models can tell us would allow for. Second, one prominent account that does speak to these ambitions relies on the false assumption that the thought experiment and the formal model are variants of the same model.

4. Lessons for Economic Methodology

While economists like Banerjee are not very clear about how thought experiment and model relate, they at least think that they are very closely related, and, importantly, that they license the same conclusion. They tend to speak of the thought experiment as an illustration of whatever it is that they show in their model. In our case this is that herd behavior is either sometimes or often caused by an informational cascade-type mechanism. Seeing that the thought experiment is about real agents, the implication is that economists think that their model establishes something about real agents, not just something about the artificial agents in their model. Especially in conjunction with introductory remarks about real financial markets or fashions, I think this shows that economists at least aspire to explain real world phenomena in a fairly direct way. This clashes with a number of accounts of what the epistemic function of economic models is.

[9] Schabas (2008) draws a similar line between mathematical models and thought experiments as I do here by describing models as 'stylized through and through', and as invoking ideal conditions such as perfect rationality.

Many economic methodologists think that economic models are not meant to tell us very much about the real world. The only thing a formal mathematical model *proves* is a conditional of the form: If all the assumptions of the model are true, then the derived conclusion must be true. Thus, if we know that all the assumptions are true in the target, then we know that the conclusion also holds for the target. Some methodologists think that this is the only thing formal models can tell us about the world (see, for instance, Aydinodat 2007). But if that was the only inference about a target system that these models allowed, they would tell us very little: It is doubtful if there has ever been an economic model for which there was a real life situation where all its assumptions held true. Hausman (1992) thinks that economic models do not directly teach us about the real world, but are only meant for 'conceptual exploration'.

Both of these accounts are called into question by the observation we made about economists' aspirations. Economic modelers seem to want to explain real world phenomena, and moreover ones where their assumptions (such as perfect rationality) need not hold true. These aspirations may of course turn out to be impossible to fulfill, and the methodologists in question may claim that their accounts are about what economic models can in fact achieve. But holding this view would commit them to an error theory that is at odds with the naturalized philosophy of science they are committed to: Economists would be systematically wrong about what it is they achieve.

Another influential account argues that economic models isolate causal capacities, which still operate in the real world, where they combine with other causal factors. The thought here is that the models tell us what would happen in an environment that is shielded from all sorts of causal factors. When these other causal factors are also present in a real world environment, they combine with the causal factor that the model isolated. Mill (1843 [1874]), 1948 [1830]) had something like this in mind when he wrote on the *method a priori*. He thought that economics studies what would result if we were all rational and motivated only by the pursuit of wealth, which is one causal force that combines mechanically with the effects of other human pursuits. Cartwright (1999) offers a similar account, but acknowledges that economics may study other causal capacities, too, and that the different causal capacities that are isolated need not combine mechanically. All we need is some ‘law of composition’.

Again, I think the thought experiments in their papers show that economists’ aspirations are at odds with this view. Contra Mill’s original version of the view, economists do not only want to isolate what people would do if they were greedy and perfectly rational. Banerjee thinks that the same mechanism is active and the same result occurs in the thought experiment, which does not make assumptions about greed and rationality, and in the formal model, which does. There is no talk of the model merely isolating a causal factor that is also present in the thought experiment. Rather, the model and the thought experiment seem to be meant to make the same claim.

One might think that, while the model doesn’t isolate what people would do if they were greedy and perfectly rational, model and thought experiment together isolate some other causal capacity. But seeing that the thought experiment already seems to explain some real world phenomena fairly directly, it is not clear what it is that could be isolated here. In any case, if what is troubling us are unrealistic assumptions about agents, we need an explanation of how a model that makes these assumptions and a thought experiment which doesn’t can both isolate the same thing.

There is one recent and influential account of the epistemic status of economic models that does speak to economists’ aspirations as we characterized them here, namely Sugden’s ‘credible worlds’ account. The main claim of this account is that we are often licensed to make inferences from models to real world phenomena because we find these models ‘credible’. In this context, Sugden discusses the stories economists tell around their models at length, including Banerjee’s restaurant story. And so his account, on the face of it, does most justice to our observations about thought experiments in the exposition of economic models.

Sugden starts from the thought that the stories economists tell around their models show that they think that we can learn from their models because there is some kind of *direct similarity* between model and target (and not just a relationship of isolation). Moreover, he thinks we can trust inferences from models more, “the greater the extent to which we can understand the relevant model as a description of how the world *could* be.” (2000, p.24) And this is what he thinks ‘credibility’ primarily consists in. Economists, in their models, construct alternative worlds which are simpler than the real world and function as analogies. Typically, the reason we take them to be good analogies is that we find them credible. Beyond the idea that we take these alternative worlds to describe how the world could actually be,[10] the only explicit thing Sugden has to say on credibility is that the alternative world should be internally coherent, and coherent with what we know about the causal factors that operate in the real world.

In Sugden’s analysis of Banerjee’s model, the restaurant story is crucial, since it is what makes the model credible, in virtue of its *familiarity*:

> By means of the restaurant story, [Banerjee] links the formal objects of the mathematical model with things in the real world with which we (the readers) are familiar. Very informally, he invites us to consult our experience of restaurants and to conclude that what is going on in his model world is in some way similar to that experience. And that is it: we are left to draw the appropriate conclusions.

That the thought experiment could play such a linking role makes sense for Banerjee because he also thinks that the restaurant story and the formal model are just variants of the same model. After having presented the restaurant story, he writes:

> Most of the rest of the paper is devoted to the analysis of a more sophisticated variant of the restaurant model.

However, we saw above that there are significant differences between the restaurant story and the formal model. Most importantly, they seem to differ in the very thing that is at stake here, namely credibility. We have seen that credibility, according to Sugden, involves appealing to what is familiar to us. But the main features that make the thought experiment familiar are not present in the formal model.

[10] In his 2009, Sugden concedes that this claim should not be taken too literally, since many assumptions in economic models could not possibly hold true in the real world, as we have seen above.

The phenomenon of an informational cascade leading to herd behavior is plausible in the thought experiment because it is cast in terms of what a real person like us might choose. We can think through the example in the way that we deliberate about our own actions. No overly idealized and unrealistic assumptions are made about what the agents know or how rational they are. The reasoning of the agents in the thought experiment is familiar, and so the causal process that is described here as a whole is credible. If that is what makes the thought experiment credible, and the formal model lacks this feature, then we are not justified to 'import' the thought experiment's credibility to the formal model. And then the 'credible worlds' account still owes us an explanation of what makes formal models credible.

Where should it take this explanation from? As I argued above, it is not obvious that we have any intuitive access to how perfectly rational agents with perfect knowledge of rationality behave, and even if we did, their reasoning patterns are not the ones that are familiar to us. They might be familiar to trained economists, because they are used to modelling in this way. But the familiarity that should be involved when we are trying to establish a relevant similarity between world and model should be a familiarity that at the very least should not depend on our having experience with the method of modelling. We do not want to end up saying that a model represents the world well because we have always modelled things in this way.

And so I think Sugden's claim that we learn from models because they are credible worlds does not go through if what he has in mind is the kind of credibility Banerjee's restaurant story has. He did not sufficiently acknowledge the fact that the restaurant story is a thought experiment, which differs from formal models in its appeal to familiar reasoning patterns - which is the main source of its 'credibility'. Thus, what should make formal models credible is still an open question.

5. Conclusions: On the Role of Thought Experiments in Economic Theory

I have argued that the informal stories economists tell before presenting a formal model deserve the status of a thought experiment. They differ from the formal models in that they employ reasoning patterns that are familiar to real people, and do not make highly idealized assumptions about rationality and knowledge. I have argued that these observations call into question a number of recent proposals on the epistemic status of economic models. Another question these observations raise is what exactly the role of thought experiments is in these papers, and what relation they bear to the formal models they are followed by.

Sugden thinks that Banerjee's restaurant story might play a kind of mediating role between the formal model and the target phenomena. I think there is some truth in that, but the story will be more complicated than Sugden suggests, and it will still leave open the question of how exactly formal models apply to the world. I want to end with some conjectures about what this mediating role might be.

First of all, I think that the restaurant thought experiment could be evidence in its own right for the claim that informational cascades sometimes or often explain herd behavior - even without any formal apparatus. Of course, the thought experiment provides defeasible evidence for this claim because a) we said the thought experimental phenomenon does not present itself as a necessity b) there is a problem of external validity: The real world circumstances may differ in relevant ways from the thought experimental setup and so we may not be able to extrapolate. But at least, in comparison with the formal model, the thought experiment has credibility in the way we characterized it above. Now if the thought experiment offers evidence in its own right, then its role in the paper may partly consist in adding independent support to the hypothesis the author is putting forward.

For economists, the thought experiment alone would not be enough, however. The number of results we can derive from it is small compared with the formal model. For instance, Banerjee uses the formal model to show that informational cascades result in Pareto inefficiency. Moreover, he can use it to study what factors exacerbate or weaken this effect. And by constructing a formal model, he contributes to a large body of economic theory that is unified by the use of perfectly rational agents. Still, the question of how the formal model relates to the world remains, since the kind of credibility that the thought experiment has cannot be appealed to.

Arguably, however, the thought experiment can still help to answer this question. It could be seen as providing an example of a more 'realistic' case that the formal model does apply to. Since the formal model seems to capture the same kind of mechanism at work in the story, it seems intuitive to view it as a model that applies to the story, and that could teach us things about the story that would not be intuitively obvious otherwise. Now if the reader is convinced that the model can be applied to and explain the story, then it is not so implausible that it could also explain a real life case of herd behavior.

Thus, I think the role thought experiments play in papers like Banerjee's is twofold: They have an independent role in supporting the hypothesis in question (that informational cascades do or can explain herd behavior), and a supportive role in providing an example of a case that the formal model does apply to.

Still, we are owed an explanation of just how the formal model applies to the story, given the differences between the two that I have highlighted in this paper - in particular the use of assumptions about rationality, and the agents' preferences and knowledge. The question of how models that appeal to some version of 'Homo oeconomicus' can apply to a world that is populated by different kinds of agents is the main question that we need to answer when we wonder about how we can learn from economic models. And this question cannot be answered by appealing to the credibility of the thought experiments that precede the presentation of a formal model: The very same question arises when we think about how the formal model relates to the thought experiment.

References:

AKERLOF, George, 2003. “Writing ‘The Market for Lemons’: A Personal and Interpretive Essay”, *Nobel Prize in Economics documents* 2001-10, http://nobelprize.org/nobel_prizes/economics/laureates/2001/akerlof-article.html (retrieved June 2015)

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

ALEXANDROVA, Anna, 2009. “When Analytic Narratives Explain”, *Journal of the Philosophy of History* 3(1), pp. 1–24.

AYDINONAT, N. Emrah, 2007. “Models, Conjectures and Exploration: An Analysis of Schelling’s Checkerboard Model of Residential Segregation”, *Journal of Economic Methodology* 14 (4), pp. 429– 454.

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

BROWN, James Robert, 1991. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*, New York, NY: Routledge.

CARTWRIGHT, Nancy, 1999. “Capacities”, in DAVIS, John, HANDS, Wade, and MÄKI, Uskali (eds.), *The Handbook of Economic Methodology,* Cheltenham: Edward Elgar.

GIBBARD, Alan, and VARIAN, Hal, 1978. “Economic Models”, *Journal of Philosophy* 75(11), pp. 664–677.

GIERE, Ronald, 2002. “Models as Parts of Distributed Cognitive Systems”, in MAGNANI, Lorenzo, and NERSESSIAN, Nancy (eds.), *Model-Based Reasoning: Science, Technology, Values*, New York: Kluwer.

GRÜNE-YANOFF, Till, and SCHWEINZER, Paul, 2008. “The Roles of Stories in Applying Game Theory”, *Journal of Economic Methodology* 15(2), pp. 131–146.

GUALA, Francesco, 2002. “Models, Simulations, and Experiments,” in MAGNANI, Lorenzo, and NERSESSIAN, Nancy (eds.), *Model-Based Reasoning: Science, Technology, Values*. New York: Kluwer.

HARTMANN, Stephan, 1999. “Models and Stories in Hadron Physics”, in MORGAN, Mary and MORRISON, Margaret (eds.), *Models as Mediators: Perspectives on Natural and Social Science*, Cambridge: Cambridge University Press.

HAUSMAN, Dan, 1992. *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.

KAHNEMAN, Daniel, and THALER, Richard H., 2006. “Anomalies: Utility Maximization and Experienced Utility” *The Journal of Economic Perspectives* 20(1), pp. 221-234.

KUORIKOSKI, Jaakko, and LEHTINEN, Aki, 2009. “Incredible Worlds, Credible Results”, *Erkenntnis* 70(1), pp. 119–131.

MÄKI, Uskali, 2005. “Models are Experiments, Experiments are Models”, *Journal of Economic Methodology* 12 (2), pp. 303-315.

MAS-COLELL, Andreu, WHINSTON, Michael D. and GREEN, Jerry R., 1995. *Microeconomic Theory*, Oxford: Oxford University Press.

MILL, John Stuart, 1843 [1874]. *A System of Logic*, New York, NY: Harper.

MILL, John Stuart, 1948 [1830]. *Essays On Some Unsettled Questions of Political Economy (Reprint)*, London: London School of Economics and Political Science.

SCHABAS, Margaret, 2008. "Hume's Monetary Thought Experiments", *Studies in History and Philosophy of Science Part A* 39 (2), pp. 161-169.

SORENSEN, Roy A., 1992. "Thought Experiments and the Epistemology of Laws" *Canadian Journal of Philosophy* 22(1), pp. 15-44.

SUGDEN, Robert, 2000. "Credible Worlds: The Status of Theoretical Models in Economics", *Journal of Economic Methodology* 7(1), pp. 1-31.

SUGDEN, Robert, 2009. "Credible Worlds, Capacities and Mechanisms", *Erkenntnis* 70 (1), pp. 3-27.

SUGDEN, Robert, 2013. "How Fictional Accounts Can Explain", *Journal of Economic Methodology* 20(3), pp. 237–243.

TOON, Adam, forthcoming. "Where is the understanding?", *Synthese*.