On the Explanation Paradox

Bachelor thesis by Annikka Lemmens

5-12-2018
Erasmus University Rotterdam
Annikka Lemmens - 376396

Contents

0: Introduction

Models are ubiquitous in economics. Models are used for prediction, decision making, and explanation in various sub-disciplines of economics. These three things often go hand in hand, for example: things that are explanatory are often predictive as well. Research issued by the European Central Bank for example mentions that prediction of macroeconomic aggregates is one of the primary functions of macroeconomic models (Amisano & Geweke, 2013). The ECB regularly bases policy decisions on predictions made using macroeconomic models. On the microeconomic level models are often said to be explanatory as well. For example, an influential model by Hansen & Wernerfelt (1989) with economic and organisational factors is said to be "as a whole successful in explaining firm profit performance" (page 405).

Even though models seem to be used extensively in the science of economics and in daily practice of economists, their use is not undisputed in the philosophy of economics. Not even the definition of a model is undisputed. Without going into the details of that debate, in this thesis when I speak about a model, I mean the following. A model is a representation of a target system. As a representation it is impossible to be a hundred per cent the same as the target system: in that case it would not be a model, it would be the target system itself. Most models in economics are highly abstract, they consist only of functions, graphs and an accompanying commentary or description of what is happening in the model.[1] It is not certain that models can represent reality accurately, and that they can give us any relevant new knowledge about the world. Does that mean that economists should not trust the models used in economics as much as they do?

In 2012 Julian Reiss published a paper called 'the Explanation Paradox'. He argued that all economic models are false but that they are nevertheless explanatory. This sparked a lot of debate. In this thesis I will explore Reiss' explanation paradox using the Spence Signalling Model as an example. Furthermore I will look at three strategies used to challenge the paradox. For each strategy I will analyse what proponents of that strategy argue, and Reiss' response. I will also give a small assessment of how convincing these arguments are to me. Furthermore, I will analyse the Spence Signalling Model again, using Robert Sugden's credible worlds approach, which is the approach that I find most fruitful. I will conclude that there is hope to get out of the explanation paradox, but for that to happen it is necessary to take a closer look at what we define as false and what we define as explanatory.

[1] However consisting only of formulas, graphs and a commentary is not a necessary for something to be an economic model. A famous counter-example is the MONIAC machine. The MONIAC uses the flow of water through pipes to represent the flow of money in the economy. (Ng & Wright, 2007)

1: The explanation paradox

In his 2012 article, Reiss explains what he defines as the Explanation Paradox. This paradox consists of the following three premises:

1) Economic models are false
2) Economic models are explanatory anyway
3) However, only true accounts can explain

Clearly, if all these premises hold, they lead to a paradox. If only true accounts can explain, and economic models are not true, it must mean that they are not explanatory. If they are explanatory then they must also be true, as models according to these premises cannot both be explanatory and false. In his article, Reiss goes into more depth on what he means with all three premises using the famous Hotelling Model as an example. I will be using another model (the Spence Signalling Model) to explain the paradox.

1.1: All models are false

According to Reiss, all economic models are false. This is not an undisputed fact, and there is even debate about whether or not models are entities that can be said to be true or false. For example Giere (1988) claimed that this was not the case. Reiss says we should read this premise as "claims and statements about models are not true in the sense of the whole truth and nothing but the truth". Models make false assumptions and they idealise. Reiss uses five classifications of different kinds of idealisations in science found in Wimsatt (2007, pages 101-102) to show in which ways models are usually wrong. In this section I will give brief description of these five problematic idealisations, in order of least problematic to most problematic. In the next section I will explain the Spence Signalling model. Afterwards I will illustrate in which ways this model is false, thereby also illustrating how these idealisations should be understood.

Of course there could be discussions about whether it is relevant for a model to be true. Here opinions differ widely. For some philosophers (like Mäki) it does matter that models are true, otherwise they could not be explanatory, which is important for him. As we will see in paragraph 3.1.1, Mäki and Cartwright therefore employ a different definition for truth than Reiss. For other philosophers (like Sugden), it does not matter that models do not contain "the whole truth and nothing but the truth". For philosophers who do not believe that models are explanatory at all, the question of whether models are truthful becomes practically irrelevant. Models are allowed to be false, as long as they fulfil their purpose, which could be to predict events, or to teach us something about modelling techniques, or to inspire us to think about the target system. Reiss sparked a lot of debate amongst philosophers of economics. Considering the fact that ideas on what the purpose of models is, and whether truth is necessary for that purpose to succeed, it seems like a good idea to analyse Reiss' paper as a starting point to think about what it means for a model to be true. As mentioned, I will discuss other conceptions of truth in paragraph 3.1.1.

1.1.1: Wimsatt's five idealisations

(1) Firstly, models can have very local applicability, even within its very limited domain. Models are always meant to just represent part of reality, but usually models only have local applicability even within their domains. This leads to a model being false if the model is applied more broadly than it was meant to be applied. A model that is meant to explain price differentiation in a duopoly setting can be false if we try to use it to explain price differentiation in a perfect competition setting.

(2) Secondly, some models are idealisations of which the conditions of applicability are never found in nature. These models leave out factors that cannot be left out in reality, or assume other conditions that are not likely to occur. In economics conditions that are often assumed are the homogeneity of consumers' tastes or a price-inelastic demand or perfect information, because these assumptions make it easier to do calculations. In reality people's tastes can differ quite a lot, the demand can depend on the price, even if it is just by a little bit and not everyone has access to all information.

(3) Thirdly, models are often incomplete, which means that they leave out other variables that have a causal relationship to the explanandum.

(4) Fourthly, models may misrepresent the interactions of the variables, perhaps showing spurious correlations or apparent independence, where dependence actually occurs.

And (5) lastly, some models may give a totally wrong-headed picture of nature. For Wimsatt this idealisation means that the model describes entities or properties, which do not exist in reality. Reiss specifies that for economics, this generally will not be a problem, as the things described are usually people or businesses, which are clearly present in reality. However, according to Reiss, a model can still be false in this fifth sense to the extent that the outcome of interest is caused by another mechanism than the one represented in the model. For example this would mean that if you make a model that predicts that owning a horse positively influences your life expectancy, this does not mean that horses give you more life years. It is more likely that the actual mechanism that is increasing the life years is the fact that rich people have the means to take good care of themselves. The mechanism postulated by the model is then completely wrong.

It is not entirely clear why Reiss chose these criteria to show that models are false. Wimsatt says these 5 options are ways in which a model can be unrealistic. This would mean that the model does not completely accurately represent reality in all its facets. Is this the same as being false? Is this the same as containing idealisations? This is not clear. Therefore, I assume that Reiss means that if a model is unrealistic in one or more of the 5 senses that Wimsatt distinguishes, it is false s. It is then not the whole truth and nothing but the truth about the target system, which means that it is not true[2].

1.2: Economic models are nevertheless explanatory

Reiss is of the opinion that all economic models are wrong in at least one of the previously discussed senses. However, he also finds that economic models are still explanatory. He mainly bases that on the fact that economists rely on models, and seem to attach value to predictions and calculations made using models. *"And perhaps more importantly, it feels explanatory" (page 48)*, is what Reiss says about the Hotelling Model.

In the 2012 paper Reiss was not that explicit about what it means for something to be explanatory. In his response to critics of the explanation paradox in 2013, Reiss mentions that this vagueness was deliberate. He says that he does not want to tell economists from an outside perspective what a genuine explanation should look like. He also mentions as a reason that accounts can be explanatory in multiple ways, and the explanatoriness of something depends on the context. It is therefore not possible to give *one* conception of what it means to be explanatory. To know what is explanatory in a certain context means that you need to take the epistemic and pragmatic purposes of the inquiry into account. Reiss also mentions that his paradox does not only hold for causal explanations, although he does reference some causal explanations in his examples to make the paradox more vivid.

Explanations seem to come in different forms, depending on the context in which they are used. There is not one definition that is universally agreed on, however, there are several separate distinctions that are often made: between ontic and epistemic explanations and between how-possibly and how-actually explanations. For some authors in the philosophy of economics only ontic, how-actually explanations are explanations. For other authors an epistemic how-actually explanation already counts. I will discuss my view on what counts as an explanation later, in paragraph 3.2.3.

[2] Considering that I already mentioned in the introduction that I do not believe that models can contain every part of the target system fully, you can already see that in my opinion models can never be the whole truth. However, I will discuss the views of other philosohpers like Cartwright and Mäki who see truth differently.

Alexandrova and Northcott invoke the distinction between ontic and epistemic explanations (2013). Ontic explanations are impersonal and objective. Causal ontic explanations place a causal structure on the world, independent of us. It means that these types of explanations tell us something about how causes in the world in fact operate, even when we would not believe in them. In contrast, epistemic explanations are subjective. An explanation is an epistemic explanation if it lessens our surprise at the outcome, making it more evident to us why an outcome obtained (Alexandrova & Northcott, 2013). According to Alexandrova and Northcott, only ontic explanations are explanations.

In response to Reiss, both Grüne-Yanoff (2013) and Mäki (2013) mention the distinction between how-possibly and how-actually explanations. A how-possibly explanation shows one particular way in which an outcome could be realised, whereas a how-actually explanation shows how one particular outcome is in fact realised. Generally, how-actually explanations are seen as explanatory. If a mechanism is at work in the real world, and this mechanism is the reason that a certain outcome is realised, most people would say that the mechanism explains the outcome. There is more debate about how-possibly explanations. If we show one way of possibly achieving an outcome, do we explain this outcome or not? Mäki would say so: "if being explanatory involves providing explanatorily relevant information about the ontic structure of the world, then no doubt correct how-possibly accounts are explanatory (and false ones are not!) (page 276)". This already brings us to the third prong of the paradox: can only true accounts explain?

1.3: Only true accounts can explain

According to Reiss "causal explanations cannot be successful unless they are true (page 43)". Stories, lies and other fictive accounts cannot explain anything when they do not contain truth, something Reiss says people feel intuitively. This is an idea that he got from Nancy Cartwright (1983). Cartwright tries to invoke this intuition using an example. Imagine a sick tree in a garden. Its leaves are falling off. We can speculate and say that the cause of this sickness is some foul water in the trunk of the tree. This would be the explanation for why the tree is sick, but it is a correct explanation only if in fact there is foul water in the trunk. If we drill a hole in the trunk, and we cannot find any water, the foul water explanation is not correct. It is still an explanation, but it is not successful at explaining.

Requiring truth for explanations has a long history in philosophy: the Deductive-Nomological model for example requires the explanandum to follow logically from the explanans, and the explanans to be true. Truth maker theory also holds that for something to be true, a truth maker is needed that makes it so. This theory is on another level than just requiring an explanation to contain truth, as it states that ontologically in the world there always needs to be something that makes a sentence true or not (Liggins, 2005). Cartwright also seems to aim for this, as she mentions that “an explanation of an effect by a cause has an existential component, not just an optional extra ingredient” (page 91).

Reiss explicitly mentions that requiring truth for explanations is the biggest downside to causal explanations. According to his paper in 2013 we should also see that the paradox holds for non-causal explanations, but he does not specifically show why those would also need to be true to be explanatory.

1.4: Summary

It seems that, according to Reiss economic models do not contain the whole truth and nothing but the truth about the target system. All models are false in at least one of the senses as discussed by Wimsatt. This reasoning is problematic, when we also state that only true accounts can explain. Intuitively, according to Reiss (and Cartwright), an explanation cannot be correct unless it contains truth. However, economic models feel explanatory. When we read about economic models, it seems like they teach us something about the world. Reiss does not make it clear what it means for something to be explanatory, according to him that is context-dependent. However, economists making models do seem to aim at explaining phenomena, the fact that they will say that a model “explains” the relationship between variables of interest points to this. Furthermore, many non-economists would say they succeed more often than not, something Reiss seems to aim at by saying “*it feels explanatory*”.

In the next section I will discuss an economic game-theoretical model. I will show how this model works and that it can be seen as strictly false in the sense of not being the whole truth and nothing but the truth. However, I still think it explains certain phenomena. In subsequent sections I will show different strategies to tackle the paradox.

2: The Spence Signalling Model

2.1: the model

Spence wants to study the informational content of data an employer has access to during the search for a new employee. This type of data includes education, job experience, race, sex and other personal observable characteristics. These characteristics can be split up into indices and signals. Indices are both characteristics that are unalterable, such as race, and characteristics that are alterable, but not by a conscious act of the person, such as age. Signals are those characteristics that are alterable by a conscious decision, such as education. In Spence's model the focus is on how the signal of education can be used on the job market.

The model Spence designed can be viewed as a game between employers and potential employees/applicants. It is a game played under uncertainty, because the employer cannot know the employee's productivity before hiring her. This means that the employer cannot know how likely it is that the employee is a highly productive worker, or a less productive worker. In the model, there are two possible outcomes of hiring someone. The employee can choose the signal she sends, and the employer will decide to hire her and what wage to offer her. An employer will have an expected marginal product of a potential employee, based on the signals and indices she possesses. Spence takes the employer to be risk neutral and he assumes that he will offer the employee exactly the marginal product he expects her to produce. Paying the marginal productivity to the worker is in line with the common practices while studying situations of perfect competition. The potential employee would possibly be willing to change her signals, if she expects to get a higher wage by doing so. However, there could be costs associated with this change of signals, called signalling costs. A prime example of signalling costs could be tuition fees or the effort necessary to change your education level. An employee will pick a signal in such a way as to maximise the difference between the potential wage and the signalling costs, so as to maximise what she would earn. Again, this is common practice: in economics it is often assumed that people want to maximise their utility and that more earnings for the same work will be beneficial to utility.

The game is played in many rounds, with the employer updating his expected marginal product of a potential employee based on the signals and indices she sends, and on the information he learned from the previous rounds. For example, if a prospective employee is 25 years old and has a master's degree, an employer could infer that this worker is probably highly productive. Supposedly, young people take fewer sick days, and a degree could be a signal of ability. If the beliefs do no longer change based on the information of the previous rounds, we call them self-confirming, which means that they are equilibrium beliefs. In the equilibrium, we always have the same wage schedules, signalling decisions, and expected marginal productivity in every new round we play, as the players do not learn anything new in each round.

Next, Spence shows the types of equilibria that could occur. The model he uses is a basic numerical model with only two types of people in the population: low-skilled people with marginal productivity of 1 (who make up part q of the population) and high-skilled people with marginal productivity of 2 (who make up part $1 - q$ of the population). The signal these potential employees can send is the level of education they take. Education level (indicated by the letter *y*) is a continuous variable in this case, which means that people can decide exactly how much they want to get of it. The costs of level *y* of education for the low productivity type are: y, for the high productivity type they are $\frac{y}{2}$. This means that getting education is more costly for low-skilled people. Spence says that these signalling costs should be interpreted quite broadly: not just in monetary amounts, but also in terms of psychic costs and time.

Spence shows that there is a possible equilibrium for a certain high enough level of education y^*. In this equilibrium low-skilled people do not get any education at all, because doing so would be too costly for them. In the model education has no value other than to signal to the employer. Any level of education below y^* does not signal competence to the employer, and is therefore useless. The high-skilled people get exactly the level $y^*$of education, to differentiate themselves from the low-skilled group. So basically, according to this model, the purpose of getting any level of education is to show that you belong to a productive group instead of an unproductive group.

What is interesting about this model is that such an equilibrium will only obtain if signalling is costly, and more costly for one group than for the other. If there are no costs attached to signalling, the signal will not be credible. This is called “cheap talk”, as anyone can make a claim about her abilities, without having something to back it up. Look at diplomas that can be bought online from websites such as “www.instantdegrees.com”, these diplomas can be bought by anyone with enough money. Such a diploma will not signal to an employer that you are highly skilled, it will only show that you had enough money to arrange a degree in 24 hours. It will therefore not credibly convey anything about your capabilities, and therefore should not provide the employer with any relevant information. It is cheap talk.

2.2: Is it false?

Following Reiss' example, I will analyse the SSM using Wimsatt's five different senses in which models can be false.

First off, the SSM could be applied too broadly. This is not a fault of the model an sich, but of the person using the model. Spence meant for his model, which in fact only describes the signalling function of education on the job market, to apply in more situations. He specifically mentions that he wants to describe all markets with many signallers who do not need to invest in acquiring signalling reputations. About the applicability of the model he mentions: "phenomena like admissions procedures, promotion in organisations, loans and consumer credit can be usefully viewed through the conceptual lens applied to the job market" (Spence, 1973, p. 356). If it turns out that in these types of markets signals do not have the differentiating function that education has in the job market, the model could be wrongly applied (Lemmens, 2017, page 2).

Secondly, the model may be an idealisation whose conditions may never be found in nature. There are some assumptions in the model which are very unlikely to be found in the real world. For example, Spence assumes that employers are risk neutral. There is a general consensus that most people are risk averse. Some scholars have tried to explain risk aversion based on expected utility maximisation(Pratt, 1964; Rabin, 2000), others based on loss aversion or mental accounting (Rabin & Thaler, 2001). Even if the reasons for risk aversion are not clear yet, risk aversion is still something we need to take into account while modelling. Most scholars these days would agree with that. Another unlikely assumption is that potential employees are either high-skilled or low-skilled. Clearly, there is not such a black and white distinction. Reality would have many more grey areas. A continuous spectrum of different skill levels seems much more plausible in reality. Even the same person could be high-skilled in one job, which leads to high productivity, but much less productive in another job. Skills are relative to the task at hand, not an absolute value that applies in all cases. The model is therefore also false in the second sense.

Thirdly, the model could leave out other relevant causal variables, which leads to it being incomplete. The SSM shows how people choose their education level, which is one outcome variable. Spence theorises that the reason for picking the signal may be to signal their productivity levels to potential employers (Lemmens, 2017 page 3). There are of course more reasons to pick a certain education level. For example it could be mandatory to go to school until you reach a certain age: in the Netherlands for example you have to be 18 years old before you are allowed stop going to school. Fear for breaking the law could be one reason to go to school until at least a certain age. Other causal factors could be that people want to raise their productivity, or perhaps they value getting more knowledge, or they enjoy the student life. These factors are not taking into account by Spence. The same goes for the other outcome variable: the hiring decision. There are other reasons for employers to make hiring decisions than just the signal about education they receive. Research shows that employers can have biases, and will then also make decisions based on the "indices" (such as gender and race), not just on the signals (Aigner & Cain, 1977). The fact that there could be other causal factors that help determine the level of education people pick, and that could influence the hiring decision means that the SSM is also false in this third sense. (Lemmens, 2017 page 3).

Fourthly, there could be a misdescription of the interactions of the variables which are included. This means that there could be spurious relationships or dependence where independence was assumed (Reiss, 2012). In the SSM it is assumed that indices and signals do not interact. This does not necessarily have to be the case. Perhaps the race of an employee is of an influence on how employers view the productivity level associated with a masters' degree. If this is the case this interaction could mean that there are many different levels of education someone would need to get to differentiate themselves from the low-skilled group, depending on his or her background (Lemmens, 2017 page 3).

Reiss adds something to Wimsatt's fourth way a model could be wrong. If a model makes assumptions that causal relationships have specific functional forms, without evidence that these modelled phenomena satisfy these specific forms, it is also wrong. A clear example in the SSM is the fact that education is assumed to be "unproductive". Getting more education does not change the productivity levels of the employees. Most people in reality do assume that the extra knowledge and skills obtained during their education increase their productivity level. An education is often seen as preparation for the job market, just look at the Dutch name: "hoger beroepsonderwijs" which translates to "higher vocational education". It is believed that following such an education will help you to perform better, when you start working. Furthermore, employers also seem to believe that education is productive, as it is not uncommon for employers to pay for their employees' classes. Having education be (too) productive would definitely change the model, as Spence himself also mentions (Spence, 1973, p. 368). Another assumption that Spence makes is that the costs of changing your signal are negatively related to your productivity. This assumption does not necessarily need to be the case for all markets Spence wants to apply his model to. For education in the job market it seems like a reasonable assumption. However, in another situation it would be necessary to study whether this assumption holds or not.

Lastly, the model could give a totally wrong-headed picture of nature. As mentioned, for Wimsatt this implies that the model describes entities or properties that do not exist. Reiss adds that for economics this fifth sense of “giving a totally wrong-headed picture of nature” can also occur if the mechanism described by the model is not the mechanism that causes the outcome. In the SSM, if people make the decision to get more education based on something completely different than wanting to signal their productivity level, this would mean that the SSM is completely wrong-headed. Intuitively there are people for whom the signalling function would not be something that is taken into consideration (or at least not after a certain level of education). For example, the reason that I am obtaining three bachelor degrees and two masters is not to show how high-skilled I am, as fewer degrees will suffice for that. I just intrinsically like learning, I am not at all trying to maximise my future earnings by adding more degrees. If this is the case for many people, then the SSM gives a very wrong-headed picture of reality to the extent that this is the case.

2.3: Is it explanatory?

As Reiss would say: the model “feels explanatory”. In my opinion, the model gives a how-possibly account of why people get education. It is plausible or conceivable that one mechanism behind getting diplomas is to show future employers how capable you are. As mentioned before, Spence suspected that similar mechanisms could be at play in markets in which there are many signallers, who do not need to invest in acquiring signalling reputations. This is for example also the case for people looking to borrow money on the credit market. The SSM does not seem to provide a how-actually explanation, at least not in all cases, as there are multiple other mechanisms that also could be at play. Moreover, it is impossible to untangle which part of the outcome was brought about by which mechanism. So whether or not this model can be seen as explanatory, it is partially going to depend on whether we think how-possibly explanations count as explanations. Grüne-Yanoff (section 3.1.1) would disagree. However, if they are explanatory, then the SSM could be an explanation of the signalling effect of education on the job market.

Does the model also provide a causal structure on the world or does it only help us to lessen our surprise at why so many people want to get a master’s degree? This is a tricky question. Alexandrova and Northcott would say that it is only the latter. Others like Mäki and Cartwright would say that it is possible that the model really provides causal structure as long as it does a good job in isolating the tendencies or capacities. Alexandrova and Northcott’s point will be discussed in section 3.2, Mäki’s and Cartwright’s in 3.1.

3: Strategies to tackle the explanation paradox

When Reiss wrote the Explanation Paradox, he was already aware of what the positions of some of his fellow philosophers would be. He included small summaries of what he thought the critics would say in his paper. Here I will clarify the three main approaches, which are to tackle each one of the premises of the paradox separately, therefore denying that there is a paradox.

3.1.1 Critics of premise one: models are not false

Prominent proponents of the fact idea that models are in fact not false are Uskali Mäki, and Nancy Cartwright. There are of course others that have argued this as well, such as Till Grüne-Yanoff and Daniel Hausman, but considering the scope of this thesis I cannot discuss every view. Cartwright originally wrote about models in physics, whereas Mäki has always focused more on economics. Despite the use of different approaches, certain aspects are very comparable.

In 2011 Mäki wrote an article exploring in what sense models are true and where this truth was located. His aim was to show that the truth of models is located inside the models, not something that is only found in claims about models. However, of more importance for the explanation paradox is his view about models as isolative representations. As mentioned before, models try to represent a target system. This target system is usually very complex, like for example the flow of money around the national economy. Models cannot show all components completely, they cannot be the whole truth. Models will isolate and keep only the important aspects of the target system (Mäki, 2011). They are also not “nothing but the truth”. To analyse the target system, models need to make idealisations to help make calculations. These idealisations are assumptions that are not strictly true about the target system. According to proponents of the idea that models are not false, the idealisations are harmless or even contribute to finding out truths.

The idealising assumptions are sometimes called Galilean assumptions, after Galileo who assumed away influences of all kinds of physical forces when conducting thought experiments. Models and (thought) experiments have certain similarities. When making a hypothesis, this hypothesis can be tested by using an experiment or by modelling what would happen in a situation that the hypothesis applies to. The Galilean assumptions are made to help the model isolate the factor that the modeller is interested in. For example, assumptions are made to neutralise other causal factors. This is the reason why economic models often assume zero transportation costs and completely identical consumer preferences. These factors can be taken out to leave just the effect of the factor of interest. The idea is that neutralising the effect of other confounding influences does not affect the factor of interest.

According to Mäki by idealising, truth bearers can be isolated. The truth bearer would be the causal mechanism at work in the model. The truth maker is the causal mechanism in the target system. Only relevant properties should be assessed for their similarity with the properties of the target system, irrelevant properties should not be treated as truth valued (Mäki, 2013). Irrelevant properties in this context would be the confounding influences, which do not affect the mechanism that you are studying. For example, if you are studying the gravitational acceleration, you are not interested in the effect of air resistance on the speed of the falling object. So you would leave air resistance out. In economics, if you are studying the effect of prices on the demand, you assume that consumer preferences are homogenous. This assumption is supposed to not influence the mechanism operating between prices and demand. Mäki then states that the assumption of homogenous consumer preferences is not true or false: it is an irrelevant property without truth value.

In this way it becomes relatively easy to get out of the explanation paradox: irrelevant properties are not true or false, and only relevant properties should be assessed. If these relevant properties are similar enough to the target system, then the model is "true". I personally think this approach is not very feasible; who decides which properties are relevant and which are not? Could we then not say of any false assumption in the model that this is not a relevant assumption, leaving only true sentences? To me it feels like the model would then still not be completely true on the whole.

Cartwright does not use the terms truth bearer and truth maker. She favours talking about capacities. "The *capacity* associated with a feature is a power that systems with that feature have to produce a specific result characteristic of the capacity" (Cartwright, 2009, page 46, emphasis in the original). Modellers use isolating tools to find out how capacities operate. To succeed in showing that a factor C has the capacity to produce effect E, the modeller should be able to argue that "(a) the specific features incorporated into the model do not interfere with C in its production of E … Beyond that, (b) these features must be detailed enough for it to be determinate whether E occurs or not; and (c) they must be simple enough so that, using accepted principles, we can derive E. Finally, and most difficult to formulate, (d) the context must be 'neutral' to the operation of C, allowing E to be displayed 'without distortion'. (Cartwright, 1998 page 45-47).

According to Cartwright (2009), economic models often do not succeed in finding capacities. The assumptions made in the model need to be either 'accepted principles' or they should be a specification of a 'neutral context'. In economics making these types of assumptions can be difficult, as there are very few acceptable principles in economics. Because there are only so few, modellers will have to include very specific structural assumptions. This lack of principles leads to the risk of overconstraint: the models have enough assumptions to isolate the capacity, but then add more assumptions beyond that. Therefore, in principle economic models could very well be true according to Cartwright, but in practice they are often not. When they are not true, they are not considered as explanatory by her.

Another important notion that is sometimes used by critics of the first premise is robustness. Results in economics are said to be robust if they obtain under different specification changes. If these specification changes are sufficiently close to realistic conditions, it could show that the model is true: it succeeded in isolating a mechanism/capacity. Kuorikoski, Lehtinen, & Marchionni (2010) compare economic robustness analysis to triangulation via independent means of determination. Triangulation is supposed to show that the same result obtains while using different methods to find it, it distinguishes the real from the illusory (Wimsatt, 1981). Economic robustness analysis could show that the conclusions do not depend on particular falsehoods and it could confirm which components of the model are most important by identifying which assumptions are crucial to the conclusions (Kuorikoski et al., 2010). People stating that models are true could use this idea by Kuorikoski to defend their claim. If a result is sufficiently robust, then the model has succeeded in isolating a mechanism/capacity. The model could then be seen as true. As Hausman (2013) stated: "either there is robustness and the true claims concerning causes and mechanisms are doing the explaining, or there is no explanation" (page 253).

In short: modellers look at the real world and try to idealise away confounding factors. This means that they are left with a model, which is just a simplified version of the real world. The model is true in the sense that it is an isolation of a true mechanism/capacity, which is something that actually operates in the real world.

3.1.2 Reiss' response

Reiss' main point of contention is about Galilean assumptions. The proponents of "models are true" say that the false assumptions made in models are irrelevant. They are Galilean assumptions meant to help isolate the mechanism/capacity. Reiss says that this is not the case. The assumptions made in economics are not the same as Galilean assumptions; they differ in at least three ways (Reiss, 2013). Galilean assumptions are usually assumed away, and are therefore not explicitly part of the model. In economic usually the assumption is an explicit part of the model. Galilean assumptions are usually about quantitative causal factors, where in economics it is not uncommon for assumptions to be categorical. In physics for example assuming away air resistance, like Galileo did, is about a quantitative factor. In economics it could for example be the case that we specify a specific type of firm. A monopolist is not just a different degree of a firm with infinite competitors. It is categorically different in other respects. The third way in which assumptions in economics are non-Galilean is that

Galilean assumptions are usually about factors which have a natural zero. Assuming away such a factor is in fact really taking out that factor: setting it to its natural occurring zero. In economics the assumptions are usually more about specifying a certain function: transportation costs are linear, utility functions have decreasing marginal utility etc. Even if some have a zero, this is not a natural zero in the same way. Zero air resistance means that air resistance does not affect the experiment. Zero price elasticity means that inelasticity influences the result. Zero is often a very specific value with a very specific effect if we take it into account. A good example of a natural zero in economics would be something like transportation costs or other type of costs, there it seems like assuming away transportation costs is not so problematic.

It seems like Reiss implies that using non-Galilean assumptions does not lead to correct isolations. The result found using these assumptions depends on all assumptions the model makes. It seems impossible to find out which subset of the assumptions drives the result (Reiss 2013).[3]

However, what about robustness? Could we overcome the problems of the non-Galilean assumptions by performing robustness checks like for example Kuorikoski et al (2010) propose? Reiss would say no. “By and large, robustness test are not possible, and if possible and performed, their result is negative”(page 52). I think that this is a hasty conclusion, especially given the fact that economists find robustness tests quite important. If a model is not robust, then usually people will start using the model less often, or only in particular situations in which the assumptions are approximately true. Finding negative results does not automatically mean that robustness checks cannot help find out if models are true. You could even say that if a negative result was found that it helps to show that the model in fact is false. This would mean that there is something to be learned about the truth value of a model by performing robustness analysis.

I find Reiss’ second argument against robustness analysis more convincing (2013). This argument starts with a short discussion of McCloskey’s A-Prime/C-Prime theorem. “For each and every set of assumptions A implying a conclusion C, there exists a set of alternative assumptions, A', arbitrarily close to A, such that A' implies an alternative conclusion, C", arbitrarily far from C.” (McCloskey, 1993, page 235). This theorem means that models are sensitive to specification changes. Reiss adds to this theorem that it would also be possible to find an A’’ arbitrarily close to A that will also result in conclusion C. This would mean that to the extent that some models are robust (so insensitive to specification changes) it could just mean that the changes are too similar to the original model. Having an A’’ that leads to the same conclusion C is not enough for Reiss, he would like a completely

[3] Here Reiss seems to agree somewhat with Cartwright. Cartwright mentioned that economic models are overconstrained and it is impossible to disentangle which assumptions drive the result.

different set of assumptions, not assumptions that are arbitrarily close. This is a common argument against robustness about the "independence of derivations". Critics of robustness analysis often state that the models that are studied during robustness analysis are not logically independent. The truth or falsity of one model will imply the truth or falsity of the other model (Orzack & Sober, 1993).

However, Kuorikoski et al. (2010) have argued that it should be the independence of individual tractability assumptions within a set of similar models that should be assessed, not the independence of models. What matters is that the tractability modelling assumptions could be thought of as not sharing the same biases and other sources of possible error. Tractability assumptions are those assumptions that are only made to help make the derivation feasible. They are made to help specify the mathematical form of the model. As long as the different but similar models use independent tractability assumptions, this could be a genuine test of robustness. It would show that the mechanism that the model found does not critically depend on the specific tractability assumptions used in the model.

3.1.3. Discussion

So, Reiss' critics say that models are in fact true in the sense that they succeed in isolating the mechanism/capacity that produces the effect. The false assumptions that are needed to derive this result are not relevant properties of the target system, and therefore do not make the model untrue. Models may not be the whole truth and nothing but the truth. However, they produce true results, which are isolations of mechanisms/capacities in the real world. I am not convinced by this approach to tackle the paradox. I find Reiss' argument that models do not isolate very plausible. The assumptions made by economic models that are useful to calculate the result cannot be said to be Galilean. They are idealisations and not just the removal of confounding causal factors. I also think that models are unrealistic and it seems like an easy way out to say that the unrealistic assumptions should just be regarded as something without truth value. Saying that only the relevant assumptions should be checked for their similarity to the target system as Mäki does, does not actually show that the model is true. It just shows that a specific part that you chose to check is similar to the target system, while not considering other assumptions that usually are still crucial to obtain the result. I do think that there is something to be learned from economic models, like I said before when discussing the SSM. I do not think that there is something to be learned from models because they are true, but that does not make them useless.

Concerning robustness analysis, personally, I believe that something genuinely can be learned from performing it. For that to happen the specification changes should be sufficiently unrelated, which I think is possible. This is what Kuorikoski et al. (2010) called independence of tractability assumptions.

However, as I said, I am not convinced that what models do is isolating by idealisation. Therefore, I do not think that robustness analysis can show that models are true in that sense. When I look at Sugden's credible worlds approach (section 3.3.1) I will come back to what robustness analysis can add.

3.2.1 Critics of premise two: economic models are not explanatory

Some philosophers like Alexandrova, Northcott and Grüne-Yanoff claim that economic models are not explanatory. This does not mean that the models are useless. They just serve other purposes instead. In his original Explanation Paradox paper Reiss mentions Hausman's account of models as conceptual explorations. However, Hausman responded to the Explanation Paradox not by claiming that models are just conceptual explorations, but by saying that they can in fact be true, which was discussed in the 3.1.1.

Alexandrova & Northcott (2013) made their view known in their paper "it's just a feeling: why economic models do not explain" as a response to Reiss. Models in economics are false, which means that according to A & N they cannot be explanatory (they agree with Reiss on premises one and three). Models fail to explain according to Kitcher's unificationist theory, according to the deductive-nomological view and according to the notion of mathematical explanation. "Mathematical explanation would require empirical confirmation precisely of the kind that that is typically absent in economic cases" (A&N page 263) Explanation according to Kitcher's idea was also discussed by Reiss (2012). However, for the current purposes it is not so relevant to explain what this view is.

The deductive-nomological version of explanation requires that the explanation includes laws, which are not present in that way in economics. In economics there could maybe be persistent tendencies instead of laws, but as those are precisely the things discovered by using the contested economic models, I do not think A&N would be impressed by the use of tendencies.

A&N believe that instead of explaining, models can help us come up with causal hypotheses that can be tested using experiments. These hypotheses can be explanatory, while the models cannot. The models are only the inspiration. A&N stress the importance of predictive and experimental success. According to them, that should be the goal of explanation: to get control over the results. Economic models usually do not give fully correct predictions: the effect in the real world is never exactly like it is in the model. Furthermore, often the predictions made by the models are qualitative and not quantitative; something that makes it hard to assess to which extent the model is explanatory. For example a model can predict that there is a positive effect, but cannot define the exact size of the effect.

As mentioned before, there is a distinction between ontic explanation and epistemic explanation. Salmon (1984) came up with this distinction. Ontic explanation consists of impersonal objective features, whereas epistemic explanation is personal. Epistemic explanations help us feel that we understand something, it lessens the surprise we feel at an outcome. A&N are of the opinion that only ontic explanation is true explanation. They say that humans like consistency and after-the-fact rationalisations. We observe something that is consistent with what comes out of the claims of an idealised model, and we assume that this means the model has explained this phenomenon.

Reiss said “more importantly, it *feels* explanatory” about the Hotelling model. This is a sentence A&N have problems with. They say that this feeling is too subjective, what can feel explanatory to one person may not feel that way to others. Also, who does it feel explanatory to? A&N say that only economists think these models feel explanatory, and that other scientists would maybe not think so. Again, the only true explanation is ontic explanation, not epistemic explanation.

Grüne-Yanoff looked at Explanation Paradox using the other aforementioned distinction of how-actually and how-possibly explanations. According to Grüne-Yanoff economic models are often not meant to just give how-actually explanations. They are just how-possibly explanations, and in that sense not explanatory. To Grüne-Yanoff how-possibly explanations are just explorations. They are meant to give a first insight into what could be the cause, which could then later on be developed into a full-fledged how-actually explanation. In this way the model could be viewed as a mere heuristic device. This heuristic model could serve to prepare hypothesis that can be tested in experiments afterwards. The final hypothesis stating the how-actually explanation will be explanatory.

Although Reiss did not want to commit to one view of what it means for something to be explanatory, he does mention that possibility hypotheses do not explain economic phenomena (2012, page 54). This seems to support Grüne-Yanoff’s idea. Grüne-Yanoff thinks Reiss is adding confusion by not specifically mentioning how-actually and how-possibly explanations, which is what got Reiss into the paradox. How-possibly explanations do not need truth: those models present possible causes without representing real world targets. How-actually models do need to be true to be explanatory. If economic models only give how-possibly explanations then they are not explanatory. This means that there is no paradox. If they give how-actually explanations but they are false then they are not explanatory either. Again this means that there is no paradox.

3.2.2 Reiss' response

Reiss recognises that models do serve other purposes as well. Sometimes they are explorations of possibilities that serve to give us new inspiration. Sometimes they serve educational purposes; they are used to teach economists (mathematical) techniques. However, that does not mean that they are not explanatory as well.

To A&N he replied that without further argument it is not immediately evident that epistemic explanations are just pseudo explanations. (Reiss, 2013) As mentioned, Reiss wanted to leave to conception of explanation open to interpretation. Economists seem to feel models are explanatory. The models serve the purposes that economists want them to. They predict events correctly for the most part, and the ideas formulated in the commentary of the models seem to make us less surprised about outcomes that we observe in the real world. Maybe epistemic explanation is enough of an explanation for the practice of economics. This seems to be a rather pragmatic account of explanation.

He also argues that economists do not think that consistency between observations in the real world with the results of highly idealised models is a sufficient condition for models to be seen as explanatory. It is a necessary condition: the model does need to predict the effect that we observe in the real world, but it is not sufficient on its own. Models need further characteristics to be explanatory. Reiss does not further specify which characteristics would be needed, but does mention that one contender is similarity to the target system (something also argued by Sugden).

To the view that models are just heuristic devices Reiss responds as follows: of course some models serve only as preparations for experiments, however it is unlikely that all models do. He argues that if models were only preparation for experiments it does not make sense for economists to use such extensive and complicated mathematical models. He argues that they should have some epistemic benefit, if they would only provide hypotheses it would be easier for economists to use other processes: random generators or crystal balls for example.

3.2.3. Discussion

Some philosophers argued that models only serve other purposes than to explain. They would mainly serve to give us inspiration to come up with real explanations. There is a lot of debate about what it means for something to be explanatory. I would personally agree with Reiss and say that it is not obvious that only ontic explanations are explanatory. I like that this seems to fit with the practice of what economists actually do. Getting the feeling that the model has led to new insights is often enough for economists, for them this seems to count as an explanation. Of course this leads to a rather subjective view of explanation, which means that something that is explanatory for me, may not be explanatory for someone else. Perhaps this ambiquity could be solved by saying that something is explanatory, if there is a consensus in the relevant scientific community that this is the case. This solution leaves room for the third prong of criticism: maybe truth is also not necessary for explanation (more on this in the next section).

What I would like about this consensus view of explanations is that this would mean that certain "explanations" which were generally trusted in the past counted as explanations at that point. They would lose the status of explanation only when people stop believing in them. There would also be downsides to such a conception of explanation. It could mean that many statements count as an explanation, and that the decision if something is explanatory becomes subjective instead of objective. It could also mean that scientific communities in different cultures or countries could have different explanations of the same phenomenon. I think this consensus view of explanation has to be worked on a lot more. Scientific consensus is concept that has been explored and is being explored by philosophers of science, so I do not want to get into it too much here.

I like the distinction Grüne-Yanoff makes between how-actually and how-possibly explanations. However, I do not automatically see why how-possibly explanations would not count as explanatory. Reiss also seems to assume this is the case, as he mentions that it is obvious that possibility hypotheses cannot be explanatory. I think how-possibly explanations can give us insights in what might go on in the real world. If we have some reason to believe that this also the mechanism that is at work in the real world, I do not see why models employing how-possibly explanations would not be considered explanatory.

Even if the model is not specific to a particular situation and a particular moment, telling us how this situation actually came about, it can still tell us about a general mechanism that could be at work. But perhaps this is already a bit stronger than how most people view how-possibly explanations. It is sometimes said that how-possibly explanations are more like conceptual explorations (Sugden, 2009): the possibility offered by the model can be interpreted as "conceivable" or "logically possible", which is a rather weak criterion. Maybe what I propose is somewhere in between how-possibly and how-actually. If we view these two as the outer points on a spectrum, what I propose is that these models give more of a "how-maybe" or "how-probably" explanation. Something that is slightly stronger than how-possibly as there is some reason for us to believe the explanation, beyond just it being conceivable, but not as strong as how-actually. We cannot say that it is definitely the case in this situation, but it is more than a mere exploration: there is some reason to believe it.

Combined with my belief that epistemic explanations count as explanations, allowing for how-possibly/how-maybe explanations would show that many more economic models could be explanatory. It would fit with the practice of economists: they build very general models that could be applicably in multiple situations and give us new insights of what could be the mechanism that produces the result that we observe in the real world. These would qualify as how-possibly/how-maybe models. This practice *feels* explanatory, even if "only" in an epistemic way. Why would we not call these models explanatory, especially considering the fact that Reiss' wanted explanatoriness to depend on the context? For economics using models in this way seems to fit very well. Of course this does not (yet) solve the Explanation Paradox if those explanations need to be true to be explanatory.

Furthermore, Reiss' claimed that models must have genuine epistemic benefit, otherwise it would be easier to think of hypotheses in another way. This does not seem like a strong argument against A&N to me. Maybe using a random hypothesis generator is just as valuable. Only after testing the randomly generated hypothesis and finding it to be true can it provide an explanation. A&N even mention that maybe models just serve to please the modeller. Maybe economists just like constructing models and feel like they can show off their prowess, whereas using a random generator does not. Just because something else is easier does not mean that modelling must have some genuine epistemic benefit. I do believe that models have epistemic benefit, and that they explain, but more because I think that epistemic how-possibly/how-maybe accounts should count as explanatory.

3.3.1 Critics of premise three: not just true accounts can be explanatory.

The approach to solve the paradox that I find most promising is to tackle the last premise. Something that is not "the whole truth and nothing but the truth" can still be explanatory.[4] Someone who explicitly stated that he did not agree with the third premise of the paradox is philosopher and economist Robert Sugden. Sugden wrote about his idea of models as credible worlds as early as 2000. Sugden specifically mentions that he works as an economist, and wanted to give a more naturalistic and pragmatic view on models. His later writings on credible worlds were inspired by Ronald Giere's book "Explaining Science" (1988).

Giere wrote about the role of models in science. Sugden interprets him as follows: "a scientific theory comprises a set of related *models* and a set of *hypotheses* linking those models with systems in the real world. A model is an abstract entity, constructed by the scientist and not making any claims about the world. Hypotheses assert similarities between the model and the world." (Sugden, 2011, page 731, emphasis in the original).

What is important about this view of models is that the model is only connected to the real world via the hypotheses about the similarities between the model and the world. This stands in sharp contrast with for example the view that models isolate a specific factor that is active in the real world. In those views the modeller starts with looking at the target system and then wonders which assumptions to make to capture just the specific factor of interest. The truth of the model would come from isolating that factor.

In contrast, Sugden's/Giere's view does not say that modellers necessarily start with the target system. What they do when building models can be more divorced from thinking about concrete properties of the real world. The model is just a construction, not a "stripped-down description of the world". Of course the real world can still be the first inspiration to come up with a model, but what happens in the model world is not connected to what happens in the target system, until similarity hypotheses are added. As the models are not connected to the target system, they can be "false", in the sense that they do not represent the target system perfectly. There is room for more

[4] Interestingly enough, Hausman mentioned "approximate truth" in his approach to the Explanation Paradox. This to me already sounds like saying models are not "the whole truth and nothing but the truth", even though Hausman tried to show that models are true in the sense that the mechanisms they demonstrate also work in the real world.

than the truth and nothing but the whole truth[5]. Models are fictional and they are their own separate world.

Alternatively, from the constructivist perspective, modelling could also come before looking at the target system. You can start with making a model, using new techniques or assumptions from other models, being inspired by different theoretical possibilities and questions, which were evoked in earlier models (Knuuttila, 2009). Afterwards, you could try to find a situation to which this model is applicable. This order is something Sugden also argues for in his paper "Explanations in search of observations". He gives a model from biology as an example that "provides a putative explanation of a regularity that has not yet been observed – but very little guidance about where it might be found" (page 725).

What Sugden argues is that by looking at models, we can still support beliefs or conjectures about properties of the real world, even if the model world is disconnected from the real world. Modellers are often not explicit about what the models exactly tell us about the world. For them it often seems sufficient to just describe whatever was going on in the model, next to what is going on in the real world, and to point out similarities between the two. For Sugden this pointed to evidence that that was what counted as explanation in the practice of economics. So an economist can create a model, describe how it works and which effects are observed within the model world. Then she will point out similar situations in the real world, and usually leaves it to the audience to draw a conclusion. This would be seen as enough to count as explanation. (Sugden, 2009)

Forming beliefs about the real world from looking at the model world follows an *abductive* inference. Abductive inference is a subcategory of inductive inference. We observe an effect in the model world, and we see the same effect present in the real world. From that we abductively infer that the mechanism that is causing the effect in the model world could also be the cause in the real world. Sugden formulated it at as follows in 2000 (page 20, emphasis in the original):

A1 – in the model world regularity R is caused by the set of causal factors F
A2 – R occurs in the real world
Therefore, there is reason to believe:
A3 – F operates in the real world

In his 2013 paper he proposes to amend the statement "therefore there is reason to believe:" to "therefore there is *some* reason to believe"

[5] This seems like good news to me, as I believe this is necessary for a model to be a model and not a copy of the target system. But again it depends on the definition of what it means for a model to be true if the whole truth and nothing but the truth is required.

Sugden does not believe that we support our beliefs about the real world by making deductive interferences, as Cartwright would like with the use of ‘accepted principles’. He also does not think inductive inferences in the form of robustness analysis can make us go from the model world to the belief about the real world. Because models are not connected to the real world, performing robustness analysis in the form of making more theoretical models with different assumptions cannot give us more certainty that the properties of the model are also present in the real world. This would just add more theoretical models that are also not connected to the real world. However, there is another role for robustness analysis as I will discuss later.

Why would we be justified to make these types of abductive inferences instead? Crucially, what matters is the notion of *similarity.* From the similarity of the effects we infer that the causes could be similar as well. According to Sugden one reason we can make these inferences if we judge the model as a similar to the real world in the sense that it ‘*could be* real’. That is “to accept that it ‘describes a state of affairs that is *credible,* given what we know (or think we know) about the general laws governing events in the world.” (Sugden, 2000, page 25, emphasis in the original). The more similar the model and the world are, the greater the degree of confidence we have that the mechanism operating in the model world is also causing the same effect in the real world. This also means that similarity is not binary: either similar or not-similar. It means that it is a matter of degree, so the confidence in our beliefs about the real worlds also goes up to a certain degree.

Similarity of a model to the target system does not make the model true, at least for Sugden. Sugden lets go of the condition that models have to be true. He could therefore still employ the idea of a model being true when it is the whole truth and nothing but the truth. This idea would then just never be realised by a model, which is fine, as truth is no longer of crucial importance for explanatoriness. If I try to incorporate Sugden’s notion of similarity in Mäki’s, Grüne-Yanoff or Cartwright’s conceptions, it could be the case that the more similar a model is to the target system, the more the mechanism that is present in the model is also present in the target system. It could then be the case that a model that is more similar, contains fewer idealisations or distorting assumptions. Of course this would not make the model “more true” as models are either true or false. But perhaps for Mäki, Grüne-Yanoff or Cartwright similarity could give an indication that the mechanism or capacity that you want to study is in fact present in the target system, which would make the model true.

Credibility is also not the same as similarity, but the more similar a model world is to the real world, the more reason we have to believe that it is credible. Other things could be contributing to similarity. There are also other things that contribute to the credibility of a model. The use of 'accepted principles' as Cartwright wants economic models to do, could be incorporated in how credible a model is. A model that makes use of more accepted principles will be more in accordance with what could be true, it becomes more credible.

Robustness analysis could be incorporated in the same way. The more robust the model result is to specification changes, the more credible it becomes that the mechanism in the model world is causing the result. Of course we still then need to take a sort of 'inductive leap', from the model world to the real world, but the more credible and similar the world is to the real world, the more justified this leap becomes. This is actually something that was brought up more by Kuorikoski and Lehtinen (2009). They thought that Sugden undervalued the epistemic role of robustness analysis in grounding model-world relationships. They are of the opinion that Sugden himself implicitly refers to robustness considerations when he makes similarity the key concept. Similarity only leads to a model being credible if the important parts of the model are to some degree similar to the modelled systems (page 129). Kuorikoski and Lehtinen claim that this means that the robustness of the important parts with respect to the auxiliary assumptions needs to be established before the comparison can take place.

Sugden mentions: "Because similarity judgments are partly subjective, what is a good explanation for one person need not be so for another. Modelling exercises that are judged worthwhile by some people may be seen as pointless by others" (Sugden, 2011, page 735). This is interesting, and especially very different from what Alexandrova and Northcott believe. More on this in section 3.3.5.

3.3.2 What does Sugden's approach mean for the SSM?

Using Sugden's approach it is possible to show that the Spence Signalling Model is explanatory, while not being true. It should be read as a fictional world, with its own rules and properties. The link between the model world and the real world should be made by the reader using abduction (so using the previously mentioned abductive inferences). In his articles Sugden often applies a textual analysis to support his claims about what modellers do in practice. This approach can be used to reinterpret what Spence has written.

Spence starts his paper by trying to pull the audience in by comparing the decision to read his paper to a decision under uncertainty in the markets he is going to analyse. He is trying to signal the importance of signalling to his audience. This is not explicitly stating that he is going to explain how such decisions are made. He mentions how the job market will be a paradigm case for such a market. However he does claim that the "conceptual lens applied to the job market" will be useful to view a considerably variety of market and quasi-market phenomena. This is something Sugden also states about Akerlof's market for lemons in his 2000 paper.

In the next section Spence starts off with a description of what a hiring decision is like. This description is speculative: “the job may take time to learn. Often specific training is required… “(page 256). This description is about properties that are observed in the real world. Next, Spence seems to move on to talking about the model world. This transition is done implicitly. He compares hiring someone to purchasing a ticket in a lottery: you do not know what you are going to get, it is a decision under uncertainty. So then he says that in the following it will be assumed that the employer will pay the certain monetary equivalent of the lottery to the individual as wage. This is clearly not an observed property of the real world anymore.

In the next parts the distinction seems a bit more muddled, but Spence seems to still be talking about the model world. Spence talks about assessment of the lottery based on the information about the signals, indices and previous information about productive capabilities of the applicants. He also mentions that “the employer’s subjective assessment of the lottery with which he is confronted is defined by these conditional probability distributions over productivity given the new data” (page 357). This is not something we directly observe in the real world. However, it is the mechanism with which the employer in the model updates his beliefs.

Next he explicitly introduces some simplifying assumptions. He does not mention whether these assumptions are realistic or not, he just states that they are made for simplicity. He does add a little example from the real world to make it believable that there are signalling costs (also in the model world). He says that education can be a signal and that we know that education is costly (in the real world). This is again an example of an implicit connection made by the modeller between the target system and the model world.

Then he introduces a critical assumption. Interestingly enough, even though it is a critical assumption, it seems to be made to make the model become interesting. “It is not difficult to see that a signal will not effectively distinguish one applicant from another, unless the costs of signalling are negatively correlated with productive capability. For if this condition fails to hold, given the offered wage schedule, everyone will invest in the signal in exactly the same way, so that they cannot be distinguished on the basis of the signal. In what follows, we shall make the assumption that signalling costs are negatively correlated with productivity” (page 358). He does not offer any anecdotes from the real world to support this assumption; it seems to be made from the perspective of the model world. Which assumption do we need to make to make this model world become an interesting world to analyse? So far, it seems to be up to the reader to see the analogue between the model world and the real world. She should supply her own reasons to think that the critical assumption is something that is credible given what we know about what goes on in the job market in reality as well.

In section 5 Spence says he will discuss the existence and properties of market signalling equilibria via a specific numerical example. In a footnote he states "Obviously, an example does not prove generality. On the other hand, if the reader will take reasonable generality on faith, the example does illustrate some essential properties of signalling equilibria" (page 361). This is a very interesting statement. He will give us an example, we ourselves need to believe that this is generalisable, and infer from that properties of signalling equilibria. Spence seems to encourage us to use inductive and analogous inference: we go from one numerical example to a general situation to another situation where it can be applicable. Do we need to stop at inferring properties of signalling equilibria in model worlds? Or do we need to take it a step further and see these essential properties as properties of the real world as well? Spence is not explicit about this, but given that he says in the introduction that the job market is just a paradigm case and he intends for his model to be applicable to a variety of markets, I would say he even wants us to take it that step further.

In sections 5 and 6 Spence analyses properties of equilibria. Quite striking again, it seems that he only analyses the model world, but there are some statements about how realistic certain procedures are peppered in.

Take this fragment for example: "A sophisticated objection to the assertion that private and social returns differ might be that, in the context of our example, the social return is not really zero. We have an information problem *in the society* and the problem of allocating the right people to the right jobs. Education, in its capacity as a signal *in the model*, is helping us to do this properly. The objection is well founded. To decide how efficient or inefficient this system is, one must consider the realistic alternatives to market sorting procedures *in the society*. But notice that even *within the confines of the market model*, there are more or less efficient ways of getting the sorting accomplished."(page 364, emphasis added).

For my purposes the meaning of this fragment is of secondary importance, what is important is that Spence clearly hints at a link between his model world and society (the real world), even though he is never explicit about how to connect the two.

Again on page 367: "we have dwelt enough upon the specifics of this model to have observed some of the effects the signalling game may have upon the allocational functioning of the market. The numerical example is not important. The potential effects and patterns of signalling are." Important for what? It seems unlikely that Spence would think the potential effects and patterns of signalling are only important in the model world or in related models. Presumably this part about patterns and potential effect hints at the fact that Spence sees his model as describing in extreme form the workings of some *tendency* which exists in the real job market, by virtue of the signalling which he claims is a property of that market. This is exactly what Sugden said about Akerlof's model (Sugden, 2000, page 5).

All in all it seems like Spence wants to show that we observe that in the real world decisions under uncertainty are often made on the basis of alterable characteristics: like in a job market we hire people based on their qualifications. Using Sugden's wording (2000, page 8): this is a regularity R that we observe in economic or social phenomena. Spence claims that this regularity can be explained by a set of causal factors F (including that signals are alterable at a cost, and a negative relationship between productivity and signalling costs), and that F tends to cause R. He supports this by making a highly formal model that is a simple, fully-described and self-contained world. He does not support his claim by using analysis of real world data. The model should make it more credible that in the real world R is also caused by F.

The credibility and similarity of the model world to the real world is implicit, but by referring to "society" and to other market which could be viewed using the same conceptual lens, Spence does seem to want us to make the inductive leap from the model to the real world. All those small references to the real world would not make sense if Spence did not intend for his highly stylised model to be applicable to real situations.

3.3.3 What does the credible worlds approach mean for the Paradox?

If we read economic models in this way, as credible fictional worlds, they do not need to be strictly true to be explanatory. As shown before, the assumptions that are necessary to make the model function are not found in the real world. This means that the model as a construct by the modeller is not a one-to-one representation of the real world. It is a simple, fully-described and self-contained world. The connection to the real world is often made implicitly by the modellers, and often it is left up to the audience, like in the example of the SSM. We go from the model world to the real world by using abduction. We observe a certain effect in the model world, where it is caused by certain factors. We also observe the same effect in the real world. If the world is similar enough to our world, we have reason to believe that the causes operating in that world are also the causes operating in our world. I believe this is quite an elegant way out of the paradox: fictional accounts can explain.

3.3.4 Critics of Sugden

Because as mentioned, I like the credible worlds approach the most, I want to look at more critics of Sugden, starting with Reiss. Sugden already wrote about credible worlds in 2000, so Julian Reiss was aware of this idea while formulating the Explanation Paradox. In his original 2012 paper Reiss interprets Sugden as saying that credibility is the same as explanatoriness. This is something Reiss disagrees with: someone judging an account to be credible not a reason to accept is as an explanation of the phenomenon per se (page 56). Reiss believes more factors are necessary for that. However, this is also something Sugden believes, although perhaps it was stated confusingly in Sugden's 2000 paper. Reiss' response in 2013 included the clarified version of Sugden's view

Credibility or plausibility is very subjective, which Reiss says can give problems for explanation (2012). Judgments of credibility are made by economists, who are influenced by their preferences about mathematical models, Just because economists have been taught to prefer highly mathematical theoretical models, does not mean that these models have to be credible. As Alexandrova & Northcott rightfully point out however, the fact that a model "feels explanatory" is also highly subjective. This is something that Reiss does not mind, but he thinks that the subjectivity of credibility judgments will lead to arbitrariness. This is something he thinks should be prevented (2013).

In the Explanation Paradox Redux, Reiss says that although it is clear now that credible is not the same as explanatory, there is still something missing from Sugdens account: a criterion which tells us if an explanation is adequate or not. The lack of such a criterion contributes to the arbitrariness of the subjective judgments of what is credible and what is not.

Another critical problem with Sugden's account is that it suffers from underdetermination. If we have multiple models that are credible and that explain the same result, we cannot say that they equally explain the phenomenon. In theory there are infinitely many models possible that could reach the same result. Not all of these models can be explanatory at the same time, for example: we cannot have too much regulation and too little at the same time. (Reiss, 2013).[6]

Reiss says that Sugden as a pragmatist should still want to give economists the tools to resolve scientific disagreements. Given the subjective nature of credibility judgments, and the fact that this

[6] The problem of underdetermination is a problem that has been discussed within the debate of scientific realism for decades, for example see for a discussion "Underdetermination of Scientific Theory" in the Stanford Encyclopedia.

account of explanation leads to underdetermination, Reiss thinks that Sugden's view is lacking something. We need something to solve the issue of underdetermination, a criterion to assert which competing explanation is better. He said that his discussion of Kitcher in the Explanation Paradox was an exploration of the role of unification in explanations. Perhaps unification could fill the role of an additional criterion.

Sugden himself does not seem to think that the subjective nature of credibility judgments and of explanation will lead to arbitrariness. He seems more pragmatic, saying that a scientific community can judge a model as explanatory if it contributes to the community's success in prediction and controlling phenomena. So long as the models serve to (help) achieve the goals set by the community, they are explanatory. This does still leave us with the question if two models using opposite causal mechanisms can still be explanatory at the same time, as long as they both help us with making predictions and controlling phenomena. It therefore does not solve the problem of underdetermination.

Reiss says that if pragmatic success is the criterion by which we judge the explanatoriness of models, then many models will not be explanatory. He gives models that failed to predict the economic crisis as an example. I, however, do not see why it would be a problem to say that those models are not explanatory for that context. The fact that models can be explanatory does not mean that every model *has* to be explanatory. Reiss also says that perhaps if this account of explanation gets developed more, it could help us perhaps get out of the explanation paradox by showing that our current models are not explanatory at all. (Reiss, 2013)

Other critics of Sugden include Grüne-Yanoff and Cartwright. Grüne-Yanoff (2009) argued that Sugden (2000) claims that we can learn too much from what Grüne-Yanoff calls minimal economic models. "Minimal models are assumed to lack any similarity, isomorphism or resemblance relation to the world, to be unconstrained by natural laws or structural identity, and not to isolate any real factors." (page 83). Grüne-Yanoff says we can only learn about impossibility hypotheses from minimal models. If we think that under no circumstance a certain effect can be brought about by a particular group of factors, and the minimal model shows one situation in which that effect does occur because of those factors, then we can say that we have lost some confidence in the impossibility hypothesis. We have therefore learnt something.

But for this to happen it must be the case that the model (1) presents a relevant possibility that (2) contradicts an impossibility hypothesis that is held with sufficiently high confidence by the potential learners. (page 97). The possibility must be relevant: we must be able to show that the differences between the model and the real world do not give reason to think that what happens in the model world could never happen in the real world. It should depict a scenario of how the world could be. The way we view the credibility of models could be like the way we view the credibility of fiction, just like Sugden (2000) also proposed. The other requirement is that the impossibility hypothesis is held with sufficiently high confidence. If the hypothesis is something that nobody believed anyway, refuting it does not teach us anything new.

Grüne-Yanoff thinks that minimal models do not support general claims about the world or about particular real-world situations (page 98), this is too big of a claim by Sugden. Going from a particular model to a general judgement about the real world depends on if we think the model is credible. This credibility judgment depends on the specifics of the model. Whether we think the models is credible or not is therefore contingent. Furthermore, he says that it seems like Sugden implies that causal principles are stable and universal, because the factors do not always operate in the real world, but sometimes they do. This would imply that we can only go from the model to particular situations, the situations in which the factors do operate, and not to a general explanation. To be able to go from the model to the particular situation, we need more than credibility of a minimal model. We would need to consider information about the real world situation as well, and we would need to compare the modelled situation to the real situation.

Personally, I think that the first part of this argument is stated rather confusingly. It seems like the only way to really get rid of the contingency would be to say that science should only be using deductive judgments, and that inductive judgments that go from a particular to a general situation should never be used. Inductive judgments going from a particular situation to a general statement are always contingent in this way. Of course it could be defended that inductive judgments have no place in science, but it does not seem very practical or realistic given the work economists actually do and given the way inductive judgments are used in all parts of life. It is like Sugden (2009) said, “Since we all find it necessary to use inductive inferences in our everyday lives, it should not be surprising to find that these are part of the practice of science too—however problematic they may for professional logicians.” (page 19). The second part of the argument I think only applies to how Sugden phrased his 2000 paper. There it seemed like credibility was the only thing that was needed for the abduction to work. This is not what he meant, as he also tried to elucidate in his later papers. Similarity of the model world to the real world is of great importance. It seems like Sugden would agree that we therefore also need to look at information about the real situation to make the judgment.

Sugden himself (2009) also mentions that he thinks that his critics, including Grüne-Yanoff, are too restrictive in which kinds of modes of inference they think are valid. They refer to “tightly restricted canons of inductive inference whose effect is to confine modellers within the structure of argument that characterises the isolation approach […] this structure of argument is too restrictive to encompass the inferences that modellers want to make, and are justified in making." (page 19). Again, it seems like Sugden wants a more practical approach, one that fits the way in which economists actually operate.

Another critic that Sugden said the same thing about is Nancy Cartwright in 2009. She interprets Sugden's credible world as "a world that contains features that occur in the real world in arrangements consistent with constraints of certain real-world institutional structures, behaving in ways dictated by principles that are at least sometimes true in those structures."(page 46). She argues that even in Sugden's account, what actually is studied is a capacity, just like she argues. [7] However, she is sceptical about how much can be learned about capacities by using the credible worlds approach.

As mentioned in section 3.1.1, Cartwright says economic models are often overconstrained which is the reason that they are false. They are supposed to be similar to Galilean experiments, but they add too many assumptions that are all relevant for deriving the result. This is exactly her problem with using abduction to go from the model world to the real world. Similar to Grüne-Yanoff she says that we have reason to believe that there are important differences between the model and the real world, and that we know that these differences matter. We know this because if we use different assumptions in the models, it will lead to different results. She does not believe that robustness analysis can save us from this problem. There is only a finite amount of assumptions that we can test. This means that robustness tests do not guarantee that the causal effect is robust to all possible combinations of assumptions. Again, she mainly seems to have a problem with the use of induction. Going from a model to a particular situation cannot be justified in her opinion.

3.3.5 Discussion

I like that Sugden says that explanation is subjective. To me this suggests that he favours the epistemic view of explanation. If explanations can be explanatory to some people but not to others, there is no reason that these explanations need to contain the whole truth and nothing but the truth. If there is greater reason to believe in a mechanism described by a model because it contains more truthful assumptions, or it is more similar to the target system, it will just be explanatory for more people. It could mean that there are perhaps gradations of explanatoriness. I think this view fits with

[7] This is similar to what Grüne-Yanoff remarked when he said that "Sugden, in claiming that one can infer general claims about the world, thus implicitly assumes the stability and universality of causal principles."

what people, including scientists, feel[8]. Some explanations are more credible than others, and maybe there could be a part of a pretty unbelievable explanation that is interesting enough to develop into a more credible explanation. This is also something Sugden mentions. (2011)

Reiss said that perhaps a new and more pragmatic/instrumentalist account of explanation based on Sugden's view could get us out of the Explanation Paradox by showing that our models are not explanatory at all. I think that this would not necessarily be the case. I think that some models will still be explanatory under this new definition, and those that are still explanatory would still be false in the strictest sense. So to get out of the Explanation Paradox we would still need to show that not just true accounts can explain, like Sugden does.

Interestingly, it seems like Sugden (2009) also seems to think about economic models as being somewhere in between how-possibly and how-actually explanations. He says he agrees with Grüne-Yanoff that how-possibly explanations are mere conceptual explorations, if we view the how-possibly explanations as just conceivability/ logically possible. He however says that for economic authors it only makes sense to include extensive discussions about the real if they "believe that their models are capable of telling us something, however speculative, about the real world." (page 23). This seems to be in between a how-actually and a how-possibly explanation.[9] The main critique of this account seems to be on the role of induction in science. As mentioned before, I do not see why induction would be problematic. We cannot deduce everything from undisputed principles, and to bring the project of science along, we will need to make statements that are not one hundred per cent guaranteed. This means that perhaps the explanations are not infallible. They came into existence using induction. If we find a situation that does not fit the explanation, perhaps the explanation needs to be revised, or the context in which the explanation is valid needs to be revised. However, this is not problematic, as I think this approach fits with the practice of science.

[8] Or alternatively, again, explanatoriness could be thought out in a scientific consensus view, in which a statement would count as an explanation if there is a consensus among scientist researching the phenomenon. Here there would be no gradation.

[9] To me it also seems that this is what Mäki aimed for when he said "If being explanatory involves providing explanatorily relevant information about the ontic structure of the world, then no doubt correct how-possibly accounts are explanatory (and false ones are not!). It looks like here he goes beyond the idea of how-possibly as just conceivable.

4: General discussion and conclusion

All things considered, there is hope to solve the explanation paradox. The most fruitful way of approaching this seems to me to be Sugden's approach of tackling the third premise. Why would only true accounts be able to explain? If fictional accounts can explain, something I think Sugden plausibly argued for, the paradox can be solved. However, it all also depends on the conception of what it means to be explanatory. If how-possibly, or "how-maybe" explanations that show how a situation could have come about show us something new about the world, then fictional models could be explanatory. Especially, if it is enough to help people feel like they got new insights into a situation, this means models can explain in an epistemic way.

I think it is convincing to say that above mentioned conception of explanation helps explain the practice of science. Scientists conceive of themselves as explaining phenomena, and they view many of their models as successful explanations. Explanation could be context depended, just like Reiss said he wanted in the Explanation Paradox Redux. Models make us feel as if we have learned something new about our world. Why could that not be enough?

I think the views that deny economic models are explanatory are too strict in what counts as explanation. Of course how-actually and ontic explanations are explanatory, but I think it needs defending to claim that how-possibly (or how-maybe) and epistemic explanations are not. The views that say economic models are true I feel are too loose in what counts as being true. Having false assumptions, but a true mechanism does not count as a true account in my opinion. That is why I prefer the looser view on explanatoriness and stricter view on falsehood that can be found in Sugden's approach.

I also feel that there maybe is not as much disagreement between the different views as is sometimes thought. Mäki for example claims that models are not false, but one way he goes about it is to claim that all auxiliary assumptions in a model, which could make it false, are irrelevant in assessing the truth value of the model. This view does not feel that different from claiming that models are false, but that should not matter for explanatoriness as long as the relevant factors are present both in the model and in the real world.

Grüne-Yanoff (2013) states that if a model correctly represents the behaviour of the factor in the target system, then the model is true. He also states that we can find out if that is the case by studying the real-world situation the model is supposed to apply to. To me this sounds very similar to what Sugden says: observe the real world situation to see if the effect is present, and observe the model to see how this effect can be brought about. Then make an abductive leap from the model to the target system. Only for Sugden this does not make the model true.

Grüne-Yanoff also says that how-possibly explanations do not need truth, as those models represent possible causes for the explanandum without representing real world targets. This again sounds similar to Sugden, only for Sugden those how-possibly explanations would probably count as explanatory, as long as we observe the same effect in the credible world as in the real world. As mentioned, I suggest that we call these types of explanations not "how-possibly" but perhaps "how-maybe" or "how-probably". These explanations offer more than just a mere logical conceivability, but they give general explanations, not guaranteed explanations for a particular explanation.

Saying that models are not explanatory like Alexandrova & Northcott do, again seems to rely on a specific notion of what it means for something to be explanatory. If feeling like you got more insight from using the model can be seen as explanatory, then A & N would also say that models could explain. They however seem to think that only ontic explanations matter. Reiss also mentioned that this needs more defending (2013).

I think the literature could greatly benefit from a more unified conception of explanatoriness and what it means for a model to be false. I think this could streamline the discussion and show that there is more unity in the different views than seems at first glance.

Bibliography

Aigner, D. J., & Cain, G. G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labour Relations Review, 30*(2), 175-187.

Alexandrova, A., & Northcott, R. (2013). It's just a feeling: why economic models do not explain. *Journal of Economic Methodology, 20*(3), 262-267.

Amisano, G., & Geweke, J. (2013). *Prediction using several macroeconomic models.* European Central Bank.

Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford, New York, United States of America: Oxford University Press.

Cartwright, N. (1998). Capacities. In J. Davis, W. Hands, & U. Mäki, *The handbook of economic methodology* (pp. 45-48). Cheltenham: Edward Elgar.

Cartwright, N. (2009). If no capacities then no credible worlds. But can models reveal capacities? *Erkenn, 70*, 45-48.

Giere, R. (1988). *Explaining science.* Chicago: University of Chicago Press.

Grüne-Yanoff, T. (2009). Learning from Minimal Economic Models. *Erkenntnis, 70*, 81-99.

Grüne-Yanoff, T. (2013). Genuineness resolved: A reply to Reiss' purported paradox. *Journal of Economic Methodology, 20*(3), 255-261.

Hansen, G. S., & Wernerfelt, B. (1989). Determinants of firm perfomance: relative importance of economic and organizational factors. *Strategic Management Journal, 10*, 399-411.

Hausman, D. M. (2013). Paradox Postponed. *Journal of Economic Methodology, 20*(3), 250-354.

Knuuttila, T. (2009). Isolating Representations Versus Credible Constructions? Economic Modelling in Theory and Practice. *Erkenntnis, 70*, 59-80.

Kuorikoski, J., & Lehtinen, A. (2009). Incredible Worlds, Credible Results. *Erkenntnis, 70*, 119-131.

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic Modelling as Robustness Analysis. *British Journal of Philosophy of Science, 61*, 541-567.

Lemmens, A. L. (2017). *The Spence Signalling Model.* Essay for the course Methodology of Economics, Erasmus Universiteit Rotterdam, Faculteit Wijsbegeerte.

Liggins, D. (2005). Truthmakers and explanation. In J. Dodd, & H. Beebee, *Truthmakers: The Contemporary Debate* (pp. 105-115). Oxford, United Kingdom: Oxford University Press.

Mäki, U. (2011). Models and the locus of their truth. *Synthese, 180*, 47-63.

Mäki, U. (2013). On a paradox of truth, or how not to obscure the issue of whtether explanatory models can be true. *Journal of Economic Methodology, 20*(3), 268-279.

McCloskey, D. N. (1993, Spring). Other Things Equal: The A-Prime/C-Prime Theorem. *Eastern Economic Journal, 19*(2), 235-238.

Ng, T., & Wright, M. (2007, December). Introducing the MONIAC: an early and innovative economic model. *Reserve Bank of New Zealand Bulletin, 70*.

Orzack, S. H., & Sober, E. (1993). A critical assessment of Levins' The Strategy of Model Building in Population Biology (1966). *Quarterly Review of Biology, 68*, 533-546.

Pratt, J. W. (1964). Risk Aversion in the Small and in the Large. *Econometrica, 32*(2), 122-136.

Rabin, M. (2000). Risk Aversion and Expected-Utility Theory: A Calibration Theorem. *Econometrica, 68*(5), 1281-1292.

Rabin, M., & Thaler, R. H. (2001). Anomalies- Risk Aversion. *Journal of Economic Perspective, 15*(1), 219-232.

Reiss, J. (2012, March). The explanation paradox. *Journal of Economic Methodology, 19*(1), 43-63.

Reiss, J. (2013). The explanation paradox redux. *Journal of Economic Methodology, 20*(3), 280-292.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton: Princeton University Press.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics, 87*(4), 355-374.

Sugden, R. (2000). Credible worlds: the status of theoretical models in economics. *Journal of Economic Methodology, 7*(1), 1-31.

Sugden, R. (2009). Credible worlds, capacities and mechanisms. *Erkenntnis, 70*, 3-27.

Sugden, R. (2011). Explanations in search of observations. *Biological Philosohpy, 26*, 717-736.

Sugden, R. (2013). How fictional accounts can explain. *Journal of Economic Methodology, 20*(3), 237-243.

Wimsatt, W. C. (1981). Robustness, Reliability and Overdetermination. In M. B. Brewer, & B. E. Collins, *Scientific Inquiry and the Social Sciences* (pp. 124-163). San Francisco: Jossey-Bass.

Wimsatt, W. C. (2007). *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Cambridge, Massachussets, United States of America: Harvard University Press.