Distortions, Models and Universality

An Analysis of Rice's Framework of Idealized Models

Atakan Dizarlar

Thesis presented for the degree of Master of Arts (MA) in Philosophy (Research)

Date of submission: July 21st, 2023

Supervisor: Prof. Dr. Jack Vromen

Advisor: Dr. Christopher Clarke

Third Reader: Prof. Dr. Fred Muller

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to Stichting Erasmus Trustfonds, Renate, Rosanne, Odette and all the other lovely people working there. Without the Erasmus Trustfonds Scholarship, studying in Erasmus University Rotterdam and writing this thesis would remain a dream. I will always be thankful to them, and the donors of the organization, for providing me with the financial opportunity to be a part of the wonderful community in Erasmus Institute for Philosophy and Economics.

I am very grateful to Jack Vromen for his valuable comments on many previous drafts and many helpful discussions. I would also like to thank him as a supportive tutor who walked me through the ups and downs of the past two years in the programme.

I would like to thank Christopher Clarke, Annalisa Costella, Moritz Hörl and Melissa Vergara Fernandez for their valuable comments on the thesis and time spared.

And last but not least, a special thanks to Fernando Varela Levy, Giacomo Sacchetti, Andrés Cano Rodriguez, Luc Lichtsteiner, Matteo Giordano, Gergana Boncheva and Jeroen de Vries. Without their friendship, emotional presence, and fascinating insights, it would be very hard to adjust to Rotterdam and complete the master's programme. I consider myself lucky to be surrounded by so many nice and interesting personalities.

Table of Contents

Introduction

One of the biggest puzzles in philosophy of science is to figure out how idealized models can provide true explanations. The sceptical argument for usage of models to study real-world phenomenon starts with these true observations:

(1) To learn[1] about real-world phenomena, scientists use models.

(2) All models are idealized.

(3) Idealizations distort certain features of the phenomena that we want to learn about.

Then, the important question is: How can we learn from a model system whose features are different from the phenomenon? Why should we care about the model's result at all?

Different philosophical accounts try to answer such worries differently.[2] For instance, a common strategy is to view scientific models as decomposable objects that can be separated into different parts in terms of their accuracy in mimicking reality and their contributions to the model result (e.g. Strevens (2009), Kuorikoski et al. (2010), Weisberg (2013), Mäki (1992; 2020)). Under such accounts, scientists first identify the *difference-making* causes or aspects of the real-world which are crucial for the phenomenon to occur. Then, the idea is to design the model in such a way that its accurate parts sufficiently resemble these difference-makers. When successful, the idealizations, which constitute the inaccurate parts of the model, only distort the *irrelevant parts* of the real-world phenomenon. So, these accounts essentially argue that models are explanatory thanks to their representational adequacy.

Against this strategy, there are accounts that emphasize the necessity of idealizations in explanation (e.g. Batterman, 2009; Batterman & Rice, 2014). These accounts argue that idealizations are so fundamental for modelling that it is not straightforward to decompose models into different parts and to separate the contributions of these parts for the model result. Particularly, Collin Rice, in a recent series of articles (Rice, 2018; 2019; 2020) and a book (Rice, 2021), tries to deliver the point that strategies which rely on accurate representation of the difference-makers do not apply to a wide range of models across different disciplines. It is quite common to have scientific models whose idealizations do distort the difference-makers because of modelling purposes. His

[1] With learning, I mainly refer to explanation and understanding of the real-world phenomena. However, until Chapter 3 where I clearly define the concepts, I will stick to broad terms like "learning about phenomenon" and "analysing phenomenon" as an aim of scientific modelling.

[2] See Aydınonat (n.d.) for a comprehensive and critical overview of the available strategies to solve this puzzle.

alternative is to think of models as idealized model systems which are fundamentally different from the real-world system they are supposed to analyse, explain, and improve our understanding of. It is called 'holistic distortion view of idealized models' (Rice, 2018). The view aims to come up with an understanding of models in which the mathematical considerations and constraints that go into model building and extracting model results are prioritized.

Since idealizations that holistically distort the model are essential and ineliminable, Rice (2018) asserts that the main aim of philosophical accounts of idealized models should be to "justify scientists' use of these holistic distortions in terms of the explanations (and understanding)" (pp. 2810). Any such justification needs to give an answer to the initial worries of how we learn from idealized models, understood as holistically distorted entities. To answer this for his holistic distortion view of idealized models, Rice (2021) first develops an account of explanation and understanding based on counterfactual relationships.

Next, Rice focuses on how holistically distorted models can be justifiably used to explain and understand. Given the counterfactual account of explanation and that idealized models can produce counterfactual information, how can idealized models provide true counterfactual information about their system? To answer this, Rice appeals to the concept of *universality*. He generalizes the notion, which is originated in theoretical physics, as "a statement of the fact that different physical systems will nonetheless display similar patterns of behaviour that are largely independent of their physical features" (Rice, 2020, pp. 832).

To summarize, there are three main elements in Rice's framework: (i) criticism of decompositional strategies and holistic distortion view of idealized models as an alternative, (ii) a counterfactual account of explanation and understanding, and (iii) universality as a particular way of justifying the use of holistically idealized models.

My primary aim in this thesis is to reconstruct the argumentative steps for all these three elements, and critically engage with them. The reason for the reconstructive purposes is due to Rice's writing style. Even though he comes up with a novel and inspiring understanding of scientific models that considers the fundamental constraints of mathematical modelling, his analysis and argumentation are intertwined with multiple case studies on models from different disciplines. This indeed helps to make the general points clear and explicit, but it also hinders the reader's ability to see the argumentative structure and small steps in the argument. Throughout the thesis, I will initially reconstruct his views or explain the underlying argumentative strategy and then demonstrate it with examples from economics. On purposes of critical engagement, the main thing I will analyse and point out its problems is the third element in Rice's framework: universality. Accordingly, Chapter 4 can be viewed as the core of the thesis.

My secondary aim is to evaluate the suitability and applicability of Rice's framework to economic modelling practices in general. The puzzlement on the relevance of scientific models for the real-world phenomena is also exacerbated by the fact that different philosophers develop their accounts in the context of a specific discipline like economics, biology, and physics. Extrapolating an existing notion or distinction from one disciplinary context and applying it to another one may be challenging as the range of modelling practices might considerably vary between the two disciplines (Zach, 2022). As Rice carves out his framework by generally analysing examples of modelling practices from physics and biology, any possible application of his framework to another discipline needs to be made carefully. To eliminate the difficulties that arise due to disciplinary context and show the applicability of Rice's account, I will heavily use models from economics in my explanation and analysis. Furthermore, I will substantiate Rice's criticism and way of perceiving models with the literature in philosophy of economics. While this thesis will show the suitability of Rice's criticism and understanding of models for economic models, it will also identify the limitations of universality-based learning from models, especially for economics.

The organization of the thesis is as follows: In Chapter 1, I introduce some of the key terms that I mention repeatedly throughout the thesis. The focus is on idealizations and their relationship with two other types of simplifications discussed in the literature: abstractions and approximations. My discussion points out that defining idealizations as distortions allows subsuming both abstractions and approximations into idealizations.

In Chapter 2, I explain and illustrate Rice's criticism of the accounts that employ the decompositional strategy. Rice rejects two out of three core assumptions of the strategy: the model decomposition assumption (D2) and the mapping assumption (D3). I also substantiate Rice's concerns for mathematical modelling by appealing to issues of integration and tractability in economic methodology. Finally, I explain Rice's alternative way of understanding models that characterize models as holistic distortions of their target systems whose use is justified by the explanations and understanding they enable that would otherwise be inaccessible.

In Chapter 3, I elaborate on Rice's counterfactual account of explanation and understanding, and how this account fits into his general framework of scientific models. To summarize, Rice argues that models explain and contribute to the scientific understanding by providing counterfactual information about contextually salient features of the analysed phenomenon. I also discuss a criticism of Rice's account collapsing into a decompositional strategy.

In Chapter 4, I introduce the concept of universality and demonstrate some drawbacks and limitations of it for model-based learning. Generalizing on various examples of models Rice analyses, I point out that appealing to universality is problematic when the model(s) at hand is targetless, and it is meant to be used generically. In such cases, it is possible to have overgeneralized and arbitrary universality classes which do not serve the purposes of justifying holistically distorted models. Similar results also arise because Rice's account allows the usage of universality without an explanation on why it occurs.

Finally, I conclude with an overview of the thesis, and a few implications resulting from my analysis of Rice's framework.

Chapter 1 – Preliminaries: Target Systems, Models, Idealizations

My aim in this chapter is to introduce the key terms that I am going to use throughout the thesis. My focus is going to be on idealizations and their relationship with two other types of simplifications discussed in the literature: abstractions and approximations.

Even though there is no general agreement on the nature, functions and epistemology of models[3], it is established that models present some simplified version or description of the real world. This suggests that many of the complexities present in the world are not included in models. As the actual phenomenon in the world has an immense level of detail, it is neither feasible nor fruitful to expect a model to be able to represent all the detail. In maybe the most influential methodological piece for economists, Friedman emphasizes how extreme this would be:

> A theory or its "assumptions" cannot possibly be thoroughly "realistic" in the immediate descriptive sense so often assigned, to this term. A completely "realistic" theory of the wheat market. would have to include not only the conditions directly underlying the supply and demand for wheat but also the kind of coins or credit instruments used to make exchanges; the personal characteristics of wheat-traders such as the color of each trader's hair and eyes, his antecedents and education, the number of members of his family, their characteristics, antecedents, and education, etc.; the kind of soil on which the wheat was grown, its physical and chemical characteristics, the weather prevailing during the growing season; the personal characteristics of the farmers growing the wheat and of the consumers who will ultimately use it; and so on indefinitely. Any attempt to move very far in achieving this kind of "'realism" is certain to render a theory utterly useless (Friedman, 1953, pp.32).

This again emphasizes that models do, and probably should, simplify many aspects of reality. There are two key things to identify based on this statement: (a) Which aspects of reality are going to be represented by models? (b) How does the simplification process will proceed? For (a), the answer is related to the notion of target systems. For (b), the main simplifying assumptions and processes I am going to analyse are idealizations, abstractions, and approximations.

1.1. Target Systems

In many accounts, including Rice's own holistic distortion view and the decompositional ones, it is accepted that instead of the real world as it is, models focus on certain aspects of it (Elliott-Graves, 2020; Suárez 2010). In a sense, these aspects are *targeted* and constitute *target systems*. Then, models try to represent these target systems. Despite the common usage of the term "target

[3] To get a sense of how severe the disagreement could get, the interested reader is advised to go through the Stanford Encyclopaedia entry of Frigg and Hartman (2020) on "Models in Science."

system", most usages do not mention what they refer to exactly (for an exception, see Elliott-Graves (2020)). But the consensus seems to be that target systems are generally viewed as parts of the world (Elliott-Graves, 2020). This naturally raises questions about how these parts are determined. My aim here is not to offer an account of target system specification, but to briefly mention what is understood by target systems in general and how they are related to simplifications, and idealizations.

Elliott-Graves (2020) mention that, firstly, any target system has a real-world context where the phenomenon occurs. Following Friedman's quote above, it could be the wheat market in the port of Rotterdam: Let us assume that wheat prices have increased in the port of Rotterdam over the last five years, and the purpose of an economist is to analyse why. Even such a statement of the fact and purpose restricts the phenomenon severely and creates a rough domain for the model. The crucial step is to identify which parts and properties of this context are *relevant* for the scientific purposes. For the purpose of our economist, some details are obviously irrelevant such as "the color of each trader's hair and eyes, his antecedents and education."

Other features of the context are not obviously irrelevant, but could be ignored safely (Elliott-Graves, 2020). The identification of such features depends on the economists and their existing knowledge about the context. For instance, the weather prevailing during the growing season is likely to have an impact on the total supply of the wheat, thus its prices. However, if the weather or the number of rainy days has been more or less stable during the growing season, then the economist can suppose that it has not been a decisive factor for the price increase due to its stability in the time period of interest. Some other relevant factors that affect the wheat prices are the oil prices as the wheat is transported and the prices of products like rice and corn that could substitute wheat. Still, knowing that these factors are relevant does not mean that they can be included in a corresponding target system. A common reason for this is that there might not be available or not decent enough quality of data to capture the factors as a part of the world that could be consistently represented by some model.

This process of choosing among all the information within the description of the real-world and determining the relevant aspects of the phenomenon and its context constitute the target specification (Jebeile, 2017). It is not a straightforward and uniform process across disciplines; there might be disagreements even within a discipline and on a particular phenomenon (Elliott-Graves, 2020). It also relies on many things including existing knowledge, theoretical components, observations, and measurements. The choices made during the target specification process is sometimes associated with *abstractions*; a concept that is highly related to the notion of idealization

(Jones, 2005; Godfrey-Smith, 2009; Weisberg, 2013). In a sense, the specification process is a process of omissions in which certain pieces of information from the real-world phenomenon is dropped. So, this set of information is abstracted away.

Although there is still much to be said on target systems and their specifications, the discussion so far captures the necessary aspects for the purpose of this thesis. To summarize, models represent some version of the real-world phenomenon. This version of the world is mostly incomplete, simplified, and abstracted and it is captured by a target system. Because of this close and indirect relationship, it is common to see that the terms such as "the target system", "the phenomenon", "the explanandum", "real world", "the real-world system" are used interchangeably. However, when we say that models are idealized (whatever it means), the reference is to the target system which are parts of the world.

1.2. Idealizations vs Abstractions

Having established some understanding of target systems, let us concentrate on idealizations. What idealizations exactly are change across different philosophical accounts; while some have a narrower sense, the others appeal to a broader understanding of idealizations. But they are generally construed as "deliberate distortion, misrepresentation and/or falsehood and amount to providing an inaccurate picture of the studied system [i.e. the target system]" (Zach, 2022, pp. 4). This way of understanding idealizations is often introduced with the notion of abstraction as the two main forms of simplifications (Levy, 2018; Zach, 2022). The issue is that it is often not clear which set of systems these notions are attributed. For instance, my brief mention of abstraction in the previous subsection concerned target systems, not the model systems. However, there are a considerable number of papers and philosophical accounts where the distinction between idealization and abstraction is imposed on model systems directly.

Let us take an example. According to the Stanford Encyclopedia entry of Frigg and Hartman (2020) on models, while "Aristotelian idealization amounts to 'stripping away', in our imagination, all properties from a concrete object that we believe are not relevant to the problem at hand, (…) Galilean idealizations are ones that involve deliberate distortions." Clearly, this understanding of Aristotelian idealizations is closely related to the abstractions, and it still mainly concerns the omissions of irrelevant features of the phenomenon. The difference is that it relates the set of abstractions to model systems directly, instead of the corresponding target system. To be clear, the result will be the same; the corresponding model would not include the abstracted feature (e.g. eye colour of wheat traders) regardless of its target. But still *where* the abstraction takes place affects the characterization of idealizations. If a philosophical account does not spend some time on introducing target systems and directly switches to the analysis of models, the classification of all simplifying assumptions about the real-world will take place within the model. In such a case, idealization-abstraction distinction is meaningful and might be useful.

Shifting abstractions into models becomes more complicated when there is an omission of *relevant* detail, feature, or factor within a model system. Note that this excludes the cases I mentioned before where a relevant detail is safely ignored and there are empirical problems regarding the measurement or observation of the relevant factor. Consider the recent increase in wheat prices in the port of Rotterdam again. For the sake of illustration, suppose that the quantity of rice traded in Rotterdam has decreased enormously in the same time span and the economist knows that rice and wheat are clear substitutes for Rotterdammers. Given this, it is still possible the economist comes up with a model to explain the price increase *without* including the market for rice. This might be simply because she wants to analyse the relationship between wheat prices and the total production of wheat. Then, is this omission an abstraction or idealization?

To answer this and understand abstractions in a more refined way, let us introduce two types of omissions: omission-as-subtraction and omission-as-extraction (Portides, 2021). The main difference between the two is that while the process of omission-as-subtraction prioritizes the features that are not essential for the modelling purposes and subtracts them from the description, the process of omission-as-extraction prioritizes the features that are deemed crucial for the modelling purposes and extracts them from the description. Clearly, there might be overlaps between the two, and both abstraction processes might lead to similar descriptions as end products, but the focus is on different set of factors.

Leaving out the eye colour of wheat traders in Rotterdam is an example of an abstraction based on omission-as-subtraction. It tries to decide which features to leave out from the complete description. Such a process implicitly assumes that "the features subtracted are necessarily known at the time the subtraction occurs, so it entails that one acts as if she possesses a list containing all the features of a particular physical system and begins to subtract in the sense of scratching off items from the list" (Portides, 2021, pp. 5883). But omission can also be interpreted as "to involve the act of extracting something and discarding the remainder" (ibid.). Such omissions-as-extractions are more about the isolation of a certain set of features of the target system from the complete set of features that the target system might exhibit. Then, such features are "abstracted away from" the target system (Portides, 2021). Accordingly, the model of the wheat without the market for rice is an example of omission-as-extraction where a relevant factor on wheat prices (i.e. the production of wheat) is isolated from the other features.

Considering the large role of models in isolating factors in some accounts, it seems like the omission-as-extraction is more suitable for model specification while omission-as-subtraction is more suitable for target specification. However, this does not change the fact omission-as-extraction is not clearly distinguishable from idealizations (Portides, 2021). Omission of rice market still misrepresents the target deliberately. Therefore, it can also be counted as an idealization.

What I aim to show with this discussion is that the general distinction between idealizations and abstractions is by no means clear. A key reason to introduce this distinction is to stress the necessity of the process of abstraction for scientific methodology. Particularly, its role in enabling the description and analysis of certain factors by isolating them in the model. As my objective in this thesis is not to develop a theory on the nature and processes of idealizations and abstractions, I will supress the distinction within model specification. In other words, I will treat abstractions as simplifying assumptions that only apply to target systems and embrace a generic notion of idealization for model systems. This also includes the omission-as-extraction type of omissions that helps certain factors to be abstracted away. The advantage of this is that it treats anything that distorts the target system, both by omitting and deliberately changing the details of an object or process, as an idealization.

1.3. Idealizations vs Approximations

Another concept that is closely related to idealizations is approximation. The notion of approximation is related to the likeness and closeness between two things. As Frigg and Hartman (2020) mention, "*A* can be called an approximation of *B* if *A* is somehow close to *B*." What counts as "close" again depends on the philosophical account. However, the content is generally mathematical closeness. Portides (2007) divides the approximation procedures into two:

> It is achieved either (a) by simplifying the relevant parts of the descriptions of individual features and properties of the physical systems in the overall theoretical descriptions, (…) or (b) by simplifying the theoretical description of the physical system as a whole in order to produce a description that is not exact but it is tractable and close enough (pp. 705).

These simplifications are mainly understood in a mathematical sense, and the appropriate notion of mathematical closeness is determined by the context. For an example of (a), consider the relationship between total savings and total investment in an economy. There is a clear positive relationship between the two in macroeconomics: When people have more money than they spend, they are likely to get an advantage of it by putting the extra amount on a savings account in a bank for an interest. As more and more savings accumulate in a bank, the bank now is able to loan out this money to potential investors on a lower borrowing interest rate. Since the investors now face a lower cost of borrowing, they can take loans to invest more easily. It is possible to model this relationship in many ways. A possible way is to think of the effect of savings to the overall investment in an economy to be a linear function of a fixed savings rate adopted by every household. This simplifies the relevant parts of the description of individual features (i.e. by supposing that every household saves at a fixed rate out of their disposable income) and simplifies the properties of the relationship between savings and investment by imposing a linear one.

For an example of (b), suppose that the aim of an economist is to determine the determinants of individual consumption. Suppose further, for the sake of illustration, he identifies three main factors that correlate with consumption: the income level of the individual, the highest level of education attained by the individual and the family size of the individual. There are multiple ways to combine all these factors and come up with a system that explains individual consumption. A linear regression assumes that all these factors that could collectively generate consumption behaviour are related in a linear form. This linear regression system may not be exact, but it is tractable, and it can be close enough to past consumption data.

There is again a potential classification issue between approximations and idealizations with mathematical content. When understood broadly, idealizations as deliberate distortions can also subsume approximations. The linear relationship between savings and investments based on a fixed parameter that captures the savings of all households can also be understood as a deliberate distortion of the target system to obey the constraints of mathematical model building. Similarly, the linearity and no-multicollinearity assumptions in a regression might also be understood as deliberate distortions for the purposes of tractability. While it is possible to keep idealizations and approximations separate (see for instance Norton (2012)), I will treat approximations as mathematical expressions of specific idealizations in the thesis.

1.4. Conclusion

Let us summarize the take-aways of the discussion in this chapter. Models simplify many aspects of reality. Which aspects of reality are going to be represented depend on their models' target systems. Even though target systems are commonly used in the philosophy of science literature, it is mostly not clear how they are formulated. The consensus seems to be that target systems are generally viewed as parts of the world, and their specification depends on various factors like scientific purposes, existing theoretical and empirical knowledge in the field and measurement techniques (Elliott-Graves, 2020). The important thing to note is that when we say that models are idealized, models idealize their target systems.

Alongside idealizations, it is possible to identify other simplification processes, mainly abstractions and approximations. While idealizations are generally understood as "deliberate distortions," abstractions refer to the omissions of actual but not immediately relevant aspects of the real world. And approximations refer to the simplifications which translate the actual factors and properties of interest into sufficiently similar theoretical objects suitable for mathematical modelling. Throughout the chapter, I demonstrated that idealizations as deliberate distortions can be interpreted in a broad manner to include both abstractions and approximations.

Subsuming both abstractions and approximations under idealizations create a broad and generic understanding of idealizations. While this can eliminate potential problems on classifying assumptions consistently, it can also lead to a discussion of assumptions which are not clear examples of idealizations under the heading of idealizations. Being aware of this, my aim is to critically engage with Rice within his own philosophical account. As Rice's aim is to focus on "the role of *distortion* in science" (Rice, 2021, pp. 4), I think this broad understanding of idealizations is also implicitly embraced by him, especially in Rice (2018; 2019; 2020). So, in the rest of the thesis, I will interpret almost all kinds of deliberate distortions as idealizations.

Chapter 2 – Rice's Criticism of Decompositional Strategies and His Alternative

After the clarificatory chapter on idealizations, I now start my analysis of Rice. In this chapter, I am going to explain Rice's criticism of the decompositional view of scientific models. Throughout the chapter, I will illustrate and substantiate his views with examples from economics and philosophy of economics.

2.1. The Decompositional Strategy

Rice argues against viewing scientific models as decomposable objects that can be separated into different parts in terms of their contributions. He mentions that the philosophical accounts that embrace the decompositional strategy assumes some version of the following set of premises (Rice, 2019, pp.181-182):

(D1) Target Decomposition Assumption: The target system is decomposable into different parts (or features). The set of parts can be further grouped as *relevant parts* (or difference-makers) and *irrelevant parts* based on their contributions to the occurrence of the phenomenon. The relevant parts are viewed as crucial for the phenomenon to occur while the irrelevant parts are viewed as insignificant aspects of the phenomenon.

(D2) Model Decomposition Assumption: The model that is supposed to represent the target system is also decomposable into different parts (or features). This set of parts can be further grouped as *accurate parts* and *inaccurate parts* based on their resemblance to the features of the target system. Moreover, the contributions of the model's accurate parts for the result can be isolated from the contributions of the inaccurate parts.

(D3) Mapping Assumption: In a successful model, the accurate parts of the model can be mapped onto the relevant parts of the target system; and the inaccurate parts of the model can be mapped onto the irrelevant parts of the target system.

The final goal of any such decompositional account is to justify the idealizations introduced in the model system. Recall the observations that we mentioned in the introduction:

(1) To learn about real-world phenomena, scientists use models.

(2) All models are idealized.

(3) Idealizations distort certain features of the phenomena that we want to learn about.

Given these true observations, it is natural to question how we can learn from a model whose features are different from its target: What ensures the relevance of the model's result for its target? To answer these questions, decompositional accounts argue and try to show that models only distort the irrelevant parts of the target system by using D1, D2 and D3. The general argument goes as follows:

D1 enables the scientist to identify the relevant parts for the target result. Mostly, these relevant parts are viewed as the difference-makers or difference-making causes for the explanandum (Rice, 2019). So, they are responsible for the target result. By D2, it is possible to separate the inaccurate parts of the model, which are the distorted parts due to idealizations, from the accurate parts. Besides, D2 also allows to isolate the accurate parts' contributions for the model result and to possibly make them the only responsible for the model result. By D3, in a successful model, the accurate parts of the model sufficiently resemble the difference-makers. Since the difference-makers are responsible for the target result and the accurate parts resemble them, only the accurate parts are responsible for the model result. Thus, even though both systems are different from each other, the impact of the differences (i.e. the distortions of idealizations) is eliminated. This is because only the inaccurate parts in the model are distorted but they distort the irrelevant parts that are not responsible for the occurrence of the phenomenon. Hence, idealizations do not get in the way in model-based explanations and learning.

2.2. An example: Functional Decompositional Approach

To give an example of such an account in philosophy of economics, let's consider Uskali Maki's functional decompositional approach, which identifies different model components and functions of these components (Maki, 1992; 2011; 2020). Accordingly, idealizations have different functions and idealizing assumptions have different contents under his account. He does not offer a specific definition of idealization but still thinks that idealizations should be distinguished from "other deformational procedures and their outcomes" (Maki, 2020, pp. 217). He underlines two main features that distinguish them: They are *deliberately false*, and their formation of "ideal" is closely connected to *perfection* (Maki, 2020). This idea of perfection sometimes takes place by setting the features or values of different parts or properties of the model to extreme, like zero or infinity. Examples include models with zero transaction costs, no frictions, perfect information in markets, infinitely large populations, or populations whose members live forever (Maki, 2020). How the other "deformational" procedures should be understood is not clear and Maki generally blends idealization as perfection and omission into a generic notion of idealization.

Maki compliments this rough description with a more extensive discussion on functions of idealizations. Their main function is to isolate certain features of the world (Maki, 2020). The real-world phenomenon is quite rich in detail and is observed in a network of complex relationships. To investigate the contribution of certain factors on the observed phenomenon, we can consider models as experiments in which idealizations are aimed to control for the other factors (Maki, 2005). It is similar to experimental controls that try to minimize the impact of factors which are not relevant for the purposes of the study but could disturb the experimental setting and results. So, modellers, by employing idealizations, omit some factors which are relevant to the behaviour of the target system or omit some characteristics of the factors that are relevant to the target's behaviour (Maki, 2011). This process of isolation is also in line with the concept of omission-as-extraction.

This account clearly employs a version of decompositional strategy, hence the name "functional decomposition." First, some set of factors that are taken to be relevant for the occurrence of the phenomenon are identified (D1). Then, this factor is mathematically represented in an economic model. In the model, the factor is isolated by idealizations to analyse its causal relationship with the phenomenon. This separates the model system into accurate and inaccurate parts since idealizations introduce distortions (D2). They mainly distort the irrelevant factors of the target, but they can also distort certain features of the relevant factor for concerns of mathematical tractability and applicability. As long as the relevant factor is isolated with *sufficient accuracy* in the model (D3), the model can provide causal mechanisms and dependency relations. Thus, idealizations are justified because of their function, and they do not affect the relevance of the model result to the understanding of its target.

2.3. Against the Decompositional Strategy

For the success of any account that follows the decompositional strategy, D1, D2 and D3 must be met. However, Rice (2018; 2019) attacks D2 (the model decomposition assumption) and D3 (mapping assumption) by giving different examples of models from physics (the ideal gas law) and biology (Hardy-Weinberg equilibrium model and optimization models). In the following two subsections, I will first summarize the general approach Rice takes while criticizing D2 and D3 respectively. Then, I will illustrate the criticism by giving examples from economics.

2.3.1 Against Model Decomposition

According to the model decomposition assumption, the model can be grouped into accurate and inaccurate (or idealized or distorted) parts and the contribution of accurate parts can be isolated from the contributions of the idealized part. Rice (2019) emphasizes that if this is indeed the case, we should be able to take out or change the idealizations in the model without affecting the contributions of the isolated accurate parts. The idea is to check what happens to the contributions of the accurate parts when the idealized part is changed or removed. If the accurate parts' contributions are not stable across different inaccurate parts, or lack thereof, how can we claim that the contributions of accurate and idealized parts are isolated? Once it is not possible to isolate the contributions of different parts, the mapping assumption becomes ineffective as well. It is no longer appealing to map the accurate parts of the model with the relevant parts of the target given that the contributions of all parts of the model are intertwined and it is revealed that the accurate parts were not the sole responsible for the model result. Hence, the argument to justify the usage of idealizations in models by saying they only distort the irrelevant parts fails.

To show that reversing the inaccurate parts of the model is generally not easy, Rice (2018; 2019) underlines the importance of mathematical frameworks in modelling. All models necessarily use idealizations to apply mathematical techniques. Without these mathematical techniques, it is not possible to get a result or an explanation from the model. Similarly, Knuuttila and Morgan (2019) mention that in the process of replacing or removing these idealizations, there are two major considerations: integration issues and tractability issues.

Integration issues are related to mathematical formulation choices which integrate the assumptions and components of the model in a specific way (e.g. mathematical moulding of Boumans (1999)). Such choices might seem like implicit features of the model, but they are central to any model building process. It is hard to change these idealizations without reformulating the model. For instance, one of the fundamental choices in macroeconometric model building is whether to model the economy with simultaneous equation systems or recursive equation systems. In simultaneous equation systems, there are some variables whose values are imposed (or simply given) on the model (i.e. exogenous variables) and some variables which are jointly dependent on one another while still being affected by exogenous variables (i.e. endogenous variables). These endogenous variables are the explanandum of the equation system. A system of equations is said to be recursive if all the endogenous variables are determined sequentially instead of jointly. The set of equations can be arranged such that the value of the first endogenous variable depends only on the exogenous variables, the second's value depends only on the first endogenous variable and the exogenous variables, and so forth.

This choice of imposing simultaneity or recursiveness to the model seems to be quite basic, but once any of them is chosen, it is not possible to switch back to the other (Knuuttila & Morgan, 2019). For instance, the theorist might believe that how people choose their investment and savings decisions can best be represented by a very complicated set of recursive equations for each individual in a model, allowing sequential determination of these endogenous variables for each person. However, if the only available data are available in aggregate terms with wide intervals (e.g. quarterly or yearly) and these data are used to specify the target, it is not possible to model people's behaviour as in the target system. Then, the theorist needs to build the model within a simultaneous framework and this assumption belongs to the idealized (inaccurate) part of the model. In such a case, the contribution of the accurate parts cannot really be tested against a different set of idealized parts. The reason is that implementing a change on the components' behaviour from being sequentially dependent to jointly dependent is not trivial and requires a complete reformulation of the model system.

One can reply that most of the distorted parts of a model is made of simpler assumptions whose impact to the overall structure of the model is severely limited compared to the above example. For instance, certain exogenous variables can be set to some *convenient* values (e.g. to zero) to smooth out the mathematical calculations and make the model system solvable. This refers to the tractability issues (Hindriks, 2006; Kuorikoski, Lehtinen & Marchionni, 2010; Knuuttila & Morgan, 2019; Mäki, 2020). In most of economic models, it is still hard to conduct the replacement of such assumptions individually because altering one might introduce different features to the system. For example, the specific value that the variable or the parameter takes might induce a functional form which affects other variables or optimal allocations.

To give an idea on this latter point, think about a common production and utility function called CES (Constant Elasticity of Substitution). The CES production function uses two factors of production capital and labour. As firms can allocate their resources on these two factors, how easy it is to substitute a factor (say capital) for labour is important. Elasticity of substitution is a ratio that gives a measure of this. As the name suggests, the CES function exhibits constant elasticity of substitution between capital and labour. But as the substitution parameter approaches to different ends, 0 and 1, the functional form varies; it turns into Cobb-Douglas production function and linear production function respectively. Thus, specific values attributed to substitution parameter have an economic meaning in the production relationships and changing them individually might not be straightforward.

To have a more particular example of economic models, let us consider the overlapping generations model. It is used heavily in growth theory to analyse the consumption-savings dynamics in an economy and the impact of redistributional policies across generations like resource transfers, education, public debt, and pension schemes (Knuuttila & Morgan, 2019; de la Croix & Michel, 2002). In the baseline case, we have two generations (young and old) alive and overlap at every period. When young, the individuals work, and they choose how much to consume or save (the savings are invested into firms). When old, they are retired, and their income comes directly from the return on their savings. As they know they are surely going to die, they spend all their income. This scenario is restricted such that it has only two types of individuals (young and old), two goods (consuming today and consuming tomorrow) and two factors of production for the firms to which the individuals invest upon with their savings (capital and labour). These are indeed idealizations that make the model tractable, and no real economy satisfies them. Note that we have not even started to impose properties on individuals' utility functions and the firm's production function to solve the model. After describing a long list of such properties, the model arrives at the conclusion that the dynamics of capital is summarized by the savings of the young that depends on their income and the rate of return on savings.

Now, we can ask the following question as Rice (2019): What would this result about the relationship between savings and capital accumulation look like without these idealizations? There are multiple ways of changing them, such as adding another generation who can work and prolonging the young period, introducing a chance factor that randomizes whether the retired will survive another period, etc. It is hard to answer this question because changing any such tractability assumption introduces additional elements to deal with: Adding another generation who can work allows the possibility of borrowing in the first young period against future wage income, but also brings additional difficulties in terms of calculation and it can be studied only in restrictive cases (i.e. logarithmic utility and Cobb-Douglas production function) (de la Croix & Michel, 2002). A stochastic chance of survival further complicates the borrowing and savings behaviour in the economy since now the retired old generation cannot simply consume all his retirement income. Thus, even before we start analysing the specific conditions that determine the behaviour of individuals and firms in the economy, it is revealed that the accurate parts of the overlapping generations model (whatever they are) can make their contributions only in this idealized, distorted and tractable framework.

These examples and the concerns of integration and tractability only scratch the surface of model decomposition assumption. To Rice (2019), the idealized parts of the model cannot be viewed as bystanders that only distort insignificant features of the model. Even though idealizations are generally disguised as seemingly innocent and economically meaningless assumptions, they are essential to both the process of model building and extracting specific model results. So, they "make a difference" (i) by formulating the mathematical structure of the model or (ii) by directly contributing to the model result. For (i), when the idealized parts are replaced, it is highly likely that a complete reformulation of the model will be necessary (e.g. simultaneity vs recursiveness). For (ii), when the idealized parts are replaced, it is highly likely that the model result will be different or not available at all (e.g. CES production function; different parameter values leading to different functions some of which give a different result, some of which do not produce any model result at all). Thus, it is not straightforward (and perhaps not even possible) to classify assumptions based on their contributions to the model result and isolate them. The target system might still have its own relevant parts or difference-makers (D1), but individual model parts cannot both accurately represent them and replicate the target result *on their own* within the model. In a sense, the role difference-maker in the target system cannot be translated to the model system.

Another factor that shows the contributions of idealizations to models is their role in the application of mathematical techniques. To demonstrate, let us take an example within game theory. Consider the Revenue Equivalence Theorem which is one of the central results in auction theory. There are multiple ways to design an auction; an auction may be first-price, second-price or all-pay. In a first-price auction, the bidder with the highest bid receives the object that is being auctioned and the auctioneer only receives this highest bid as the price. While in the second-price auction, still the bidder with the highest bid receives the object but this time, only the bidder with the second highest bid must pay her bid to the auctioneer. In all-pay auction, the highest bidder gets the object, but all bidders must pay their bid. The rules of the auction affect the bidding strategies of the participants. Surely, my strategy when I am paying my bid regardless of the others' bid will not be the same as the case where I can submit the highest bid but get away with paying the second-highest bid.

How do these different types of auctions compare from the perspective of the auctioneer? Since the auctioneer is interested in the expected sales price (or his revenue), his aim is to come up with the auction that generates the highest revenue. Unfortunately for the auctioneer, Revenue Equivalence Theorem states that all aforementioned auction types generate the same expected revenue:

Revenue Equivalence Theorem: *Let us focus on auctions that assign the object to the highest bidder. Suppose the bidders have independent and identically distributed valuations and are risk neutral. Then, any symmetric and increasing equilibrium of any auction such that the expected payment of bidder with value 0 is 0, yields the same expected revenue for the auctioneer.*

Many auction models in equilibrium award the object to the bidder with the highest value, i.e., a bidder's probability of winning is just the probability that her valuation is the highest. Similarly, in all these models the bidder with no valuation for the object has no chance to profit, so, the theorem says, all these mechanisms will yield the same economic result. The result of course depends on idealizing assumptions. To describe them:

(i) Each bidder has a valuation which is distributed according to a continuous probability function F(.). Since each player can have infinitely many possible valuations, any bidder's valuation is treated as a random variable and the probability of a bidder's valuation is described in ranges of numbers.

(ii) Each bidder's valuation is independently distributed; meaning that knowing the actual value of a bidder's valuation does not impact the probability distribution of another bidder's valuation.

(iii) Each bidder's valuation is identically distributed; meaning that each valuation has the same probability distribution as others.

(iv) Each bidder is risk neutral as such they are indifferent between a sure amount M and any gamble whose expected return is M.

(v) The focus of analysis is only on the case where each bidder forms a strategy based on their valuation and each of these strategies are increasing (the higher your valuation, the higher your bid is) and symmetric (everyone employs the same function as their strategy).

These idealizing assumptions within the Revenue Equivalence Theorem do not simply note certain features of real bidders' valuations, strategies and risk-attitudes are irrelevant; instead, they "pervasively distort" the behaviour and interactions of real bidders to "allow for the application of mathematical tools" (Rice, 2019, pp. 191-192). Without the assumptions on the distribution of individual valuations —crucially, the independent and identical distribution assumption— it is not possible to employ statistical modelling techniques and conduct the analysis based on winning probability. Similarly, the assumptions on the strategies of the bidders allow us to analyse how the bidders are going to behave in the equilibrium given their valuations. Changing these idealizations such as breaking the symmetric strategies assumption, would lead to a case in which two bidders who have the same valuation could act differently. This would complicate the model and the whole analysis greatly; the model might not be even solvable. Since these idealizations are fundamental for applying mathematical techniques and reaching to a result from the model, they are pervasive, and it is not possible to eliminate them from the explanations provided by the model (Rice, 2019). Hence, we cannot "isolate the contributions made by some accurate part(s) of the model from the contributions made by these idealized parts" (Rice, 2019, pp. 192).

2.3.2 Against Mapping Model Parts to Target Parts

Now, let us continue with the criticism of D3. According to the mapping assumption (D3), the accurate parts of the model can be mapped onto the relevant parts of the target system; and the inaccurate parts of the model can be mapped onto the irrelevant parts of the target system. The previous argument already showed that mapping the accurate parts to the relevant parts in terms of contributions is generally not possible: The accurate parts (or any part) cannot reproduce the target result on their own; they can make their contributions only when they are combined with idealizations that are necessary for formulating the model environment and making it tractable, and for applying certain mathematical techniques without which there might be no model result.

In this subsection, the focus is going to be on representation. Rice (2018; 2019; 2020) stresses that the accuracy of model representation of the relevant parts is questionable. Particularly, Rice (2019) mentions that idealizations often distort the relevant parts of their targets as well. So, even though one could separate the model system into accurate and inaccurate parts and isolate their contributions, the accurate parts alter the features of relevant parts. Then, the mapping assumption fails and the general argument of decompositional accounts breaks down: Since these relevant parts are viewed as the difference-makers for the target phenomenon and are responsible for the explanandum, now some of the inaccurate parts are also responsible for the result.

The reason for why the relevant parts are often distorted in scientific modelling is again related to the constraints of mathematical modelling. To investigate the difference-making factors of the target system within the model, the scientist somehow needs to include them in the mathematical framework. It mostly leads to the simplification or removal of certain aspects of those factors that known to make a difference for the target system. This is because in order for the mathematical model to be integrated, tractable and solvable, it needs to limit its attention only to certain features of the target. Otherwise, certain mathematical techniques might not be applicable anymore or the underlying mathematical system of the model might turn out to be overdetermined and result in no solution. To avoid this, the scientists frequently utilize idealizations which distort some of the relevant features and actual processes of the target system.

For an example of this, let us consider first perfectly competitive markets in economic theory. A perfectly competitive market describes a market in which every seller and every buyer takes prices as given. In other words, both sellers and buyers take prices as parameters, rather than choice variables. Going back to the wheat example, there are many wheat farmers over the world. Their number is so high that no single seller who grows and harvests wheat has any control over the price; all sells it at the market price. Of course, they do not have to sell it if they consider the market price as too low. But they cannot bargain or force for a higher price as there are many other wheat producers over the world who could supply almost the same product. Thus, these sellers (or firms) are price takers, not price makers. The analytical translation of this idea into a mathematical model is to take prices as given parameters (Hands, 2016). Buyers of wheat are also in a similar position. Given a market price, they can only decide to take it or leave it. Such a market is said to be perfectly competitive.

A perfectly competitive market is not a complete model by itself, but it has created the basis of many mathematical models in economic theory such as the Walrasian general equilibrium model. It is a highly idealized representation of a market and there is a long list of highly restrictive conditions for it, like free entry and exit to the market, market participants to be completely informed, negligible search costs, etc. No real-world market can satisfy these requirements simultaneously. Hence, any model with a perfectly competitive market to analyse equilibrium behaviour of market participants is a model whose idealizations also distort the relevant parts of the target phenomenon. If an economist would use the perfect competition assumption in a model to analyse the market price of wheat in the port of Rotterdam, he would be distorting the difference-making features of the target. These are difference-making features because what kind of competition the market exhibits has a direct and crucial influence on the prices both in the real-life and model settings. The competition could also be modelled in a way that a certain producer is the monopoly over the wheat production, or the interaction between a seller and producer can also be modelled in a bargaining setup where they can alternate offers (à la Rubinstein (1982)).

To restate the core idea, idealizations like perfect competition simplify the modelling process and the mathematical analysis substantially. For instance, thanks to distortions of perfect competition, the existence and the stability of market equilibrium can be ensured, and this allows the economists to come up with counterfactual analysis by varying parameter values at the equilibrium. However, this is not in line with the decompositional strategy. The reason is once the relevant parts of the target system are altered within the model, one can no longer map these relevant parts only to the accurate parts of the model. So, the mapping assumption (D3) fails. Rice (2018; 2019) emphasizes that the idealization of relevant parts is widespread in several other disciplines like biology, chemistry, and physics. Therefore, given that the aim of decompositional strategies is to show that it is the non-idealized parts of the model that "do the real work," the decompositional strategy will frequently fail.

2.4. Rice’s Alternative: The Holistic Distortion View of Models

After rejecting the decomposition strategy in which models are viewed as systems some of whose parts are idealized, Rice offers an alternative way of understanding models. It is called ‘holistic distortion view of idealized models’ (Rice, 2018). Its fundamentals have already been hinted in his criticism, and accordingly, the holistic view is going to be concentrated on the mathematical considerations and constraints that go into model building.

Rice’s alternative is to think of models as idealized systems which are fundamentally different from their corresponding target system, and thanks to which the scientists are able to analyse, explain and understand the real-world phenomenon (Rice, 2019). This account is practically motivated and aims to capture how the scientific practice produces explanations:

> Many (if most) of the idealized models that are used to explain in science holistically distort the entities, processes, and difference-making features of their target system(s) in order to allow scientific modellers to utilize various mathematical modelling techniques that would not otherwise be applicable. Applying these mathematical modelling techniques, in turn, allows these modelers to access scientific explanations that would not otherwise be accessible (Rice, 2019, pp. 2809).

So, the core of this account has two elements. The first is that in a wide range of cases, idealized models pervasively distort the fundamental nature of the entities, processes, and features of their target systems—including both relevant and irrelevant features. So, their distortions are holistic. Second, these idealizing assumptions often move scientists to an entirely different representational framework, in which the mathematical tools necessary to explain and understand the phenomenon of interest are applicable. These different mathematical modelling frameworks represent different features and patterns of the system in different ways and allow for the use of different techniques for deriving the behaviour of the system. As a result, these idealizations are often necessary for the models to provide a particular explanation (or understanding) of the target phenomenon. A model without these idealizing assumptions would be unable to use the mathematical modelling techniques that are required for extracting the explanatory information of interest.

The characterization of models as holistically distorted entities clearly recognizes the limitations of scientific practice and prioritizes the mathematical considerations that go into model building. It is not always necessarily the case that models distort *all* features and processes of their target systems. Instead, it emphasizes that we, as philosophers, are rarely able to identify different parts of a model and demarcate its accurate parts from the inaccurate, idealized parts. Accordingly, Rice (2018) asserts that the main aim of philosophical accounts of idealized models should be to “justify scientists’ use of these holistic distortions in terms of the explanations (and understanding) they enable them to achieve that would otherwise be unattainable” (pp. 2810).

2.5. Conclusion

In conclusion, Rice mentions that the decompositional strategy is composed of three core assumptions: D1, D2 and D3. Maki's functional decompositional approach is an example of such a strategy. Rice attacks the accounts that employ the decompositional strategy by criticizing D2 and D3. In this chapter, I explained his criticism of D2 and D3 by using examples of idealizations and models from economic theory. I also substantiated Rice's concerns for mathematical model building by appealing to issues of integration and tractability in economic methodology. The necessity of idealizations to account for the considerations of integration and tractability in the process of model building is a highly valid, and sometimes underestimated, point on economic models.

For the model decomposition assumption (D2), Rice mentions that if the models were indeed decomposable, we would be able to remove or change the idealizations and their content without influencing the contributions of the accurate parts to the model result. The problem is that removing or changing idealizations without touching the model result is not the case most of the time. This is because all models necessarily use idealizations to apply mathematical techniques and get model results. So, even though idealizations in the inaccurate parts might be disguised as inessential, meaningless and innocent, they are fundamental for extracting model results. Therefore, it is not straightforward (and possibly not even possible) to classify assumptions based on their contributions to the model result and isolate them.

According to the mapping assumption (D3), the accurate parts of the model can be mapped onto the relevant parts of the target system; and the inaccurate parts of the model can be mapped onto the irrelevant parts of the target system. Rice (2019) states that idealizations often distort not only the irrelevant parts of the target, but also the relevant parts. Then, although one could divide the model system into accurate and inaccurate parts and isolate their contributions, the accurate parts still typically represent the relevant parts in an idealized way. Thus, it is not possible to map the accurate parts with the relevant parts as some of the inaccurate parts are also responsible for the result. The distortions of relevant parts are also introduced for the mathematical model to be integrated, tractable and solvable.

After rejecting the decomposition strategy, Rice offers an alternative way of understanding models called 'holistic distortion view of idealized models' (Rice, 2018). It suggests, instead of trying to decompose models into an accurate representation of relevant features and the distortion of irrelevant features, idealized models ought to be characterized as holistic distortions of their target systems whose use is justified by the explanations and understanding they enable that would otherwise be inaccessible. So, the account views idealizations as tools to get model results. I think this way of looking at idealizations captures the general practice of economic modelling well and I demonstrated its suitability with various examples from subfields like economic growth and game theory.

Chapter 3: How Idealized Models Explain and Provide Understanding about their Targets?

In the previous chapter, we mentioned that Rice embraces a holistic view of models in which models are viewed as holistically idealized entities, instead of being composed of piecemeal idealizations affecting only certain parts. Under this framework, Rice is back to the start and still needs to answer to the question of "How do models, as holistically distorted entities, provide learning about the real world?" In the current and the next chapters, I am going to concentrate on his answer to this.

His answer appeals to the concept of universality and universality classes. Before we go into the details of universality, we need to understand what kind of information models can provide according to Rice. Specifically, Rice argues that models explain and contribute to the scientific understanding by providing counterfactual information about contextually salient features of the analysed phenomenon. This counterfactual information can be both about counterfactual dependence and independence. Even though successful instances of both explanation and understanding require modal information, Rice separates the two concepts. I will first talk about Rice's account of explanation in Section 3.1, and then switch to understanding in Section 3.2.

3.1. Rice's Account of Explanation

3.1.1. Unifying Causal and Noncausal Explanations with Counterfactuals

Rice (2021) underlines the counterfactual (or modal) information that models produce for purposes of explanation. He states his account of explanation as follows (Rice, 2021, pp. 108):

> *The Counterfactual Account of Explanation:* In order to explain phenomenon P, an explanans E must include a set of true modal information about P regarding both how the contextually salient features on which P counterfactually depends account for the occurrence of P and show that the contextually salient features on which P does not counterfactually depend are irrelevant to P.

So, the key element in providing an explanation is to offer counterfactual dependence and independence information on contextually salient features of the explanandum. I admit that I struggle to present and explain this account in the same manner as Rice. While this is partly due to my limited knowledge and study on theories of explanation, it is also due to Rice's (2021) exposition of the account; it seems compressed and is intertwined with many case studies and other theories of explanation. It is also hard to see how this account is supposed to work in general, and how it fits with his understanding of models. To remedy these issues, I will first give my own reconstruction of his account of explanation step-by-step in a modelling scene. Then, I will go through each step and expand on the crucial and non-obvious concepts they utilize:

(1) Before the practice of modelling, the scientist identifies contextually salient features of the explanandum. This identification process depends on the scientist's aims, existing theories, limitations on measurement and observation of the target phenomenon, etc.

(2) The explanandum and its contextually salient features are represented in the holistically distorted models that use idealizations as tools to extract model results.

(3) The initial set of contextually salient features are captured by the model. It is highly likely that these features now are distorted. Based on the relationship between these features and the model result, we can separate this set of features into (a) the ones that the target phenomenon counterfactually depends on and (b) the ones that the target phenomenon is counterfactually independent of.

(4) We say that the model explains the explanandum when it provides a set of true counterfactual information on (a) and (b). Particularly for (b), the idea is to show why those supposedly contextually salient features turned out to be irrelevant.

On (1)-Contextually salient features: Since a successful explanation must produce a set of true counterfactual information about the explanandum's dependence and independence on contextually salient features, it is important to talk about how the contextually salient features are identified. Note that even though Rice criticizes model decomposition assumption (D2), he seems to embrace (though not explicitly) the target decomposition assumption (D1). Accordingly, the target system is decomposable into different parts, or features. In determining which features are salient or non-salient to providing an explanation, Rice (2021) highlights the role of the context where the explanation is pursued. Similar to Potochnik (2017) and Woodward (2003), Rice (2021, pp. 102) states that "the pragmatic features of the explanatory context will determine the set of features whose counterfactual relevance and irrelevance need to be accounted for in providing the explanation."

I find this idea to be closely connected to the target system specification. Recall that target system specification is the process of choosing among all the information within the description of the phenomenon and determining its relevant aspects. It is not a straightforward and uniform process across and within disciplines. As I mentioned in Chapter 1, target system specification is not an area studied as commonly as model system specification, and it is kind of a black box between the real-world explanandum and the model. So, unsurprisingly, Rice (2018; 2019; 2020; 2021) does not present a detailed account on how this specification works. However, Rice (2021) mentions some of the factors that shape the identification process. These include existing knowledge and modelling traditions in the field, how the phenomenon is observed or measured, which factors are deemed important for the purpose of the model, etc.

On (2)-Holistically distorted models: This is basically the holistically distorted view of models that we analysed extensively in the previous chapter.

On (3)-Grouping of contextually salient features: Now, take the inital set of contextually salient features. By assessing them inside the model and based on its results, we can separate this set of features into (a) the ones that the target phenomenon counterfactually depends on and (b) the ones that the target phenomenon is counterfactually independent of. Separating them into (a) and (b) implicitly assumes some version of model decomposition assumption (D2). This is because the features represented in the model are now separated, not those of the target system. But this one is different from the one embraced by decompositional strategies in that (i) it is not based on accurate representation, (ii) it does not state the contributions of (a) and (b) to a particular model result can be isolated from each other. I will come back to this point on whether Rice's account collapses into a decompositional strategy in Section 3.1.2.

On (4)-Counterfactual information: Let us expand upon the counterfactual information produced by the model. Rice follows Woodward's criterion stating, "explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways" (Woodward, 2003, pp. 11; as cited in Rice (2020)). However, Rice also puts great emphasis on counterfactual independence relations:

> Unfortunately, most accounts of explanation focus exclusively on the features that make a difference and, therefore, miss the importance of actually showing that certain features are irrelevant to the occurrence of the explanandum. Both kinds of information are often required to provide the desired explanation, and neither should be universally privileged as more important (Rice, 2021, pp. 99).

I believe that the underlying idea is when a scientific theory hints that a feature is contextually salient at the stage (1), it is supposed that it is going to be relevant for the model result. So, when it turns out that the feature is not relevant, the explanation of the phenomenon needs to account for it. To Rice (2021) many scientific explanations are built upon such knowledge of why specific features are counterfactually irrelevant to the occurrence of the explanandum. As an example, recall the statistical modelling in auctions and the Revenue Equivalence Theorem that we analysed in the previous chapter. It can show most of the details that can potentially contribute to how much the winner is going to pay is actually irrelevant from the perspective of the seller as long as the highest bidder wins.

However, this does not mean that *all* explanations are going to account for irrelevant features. Rice (2021) expresses that it is possible to have explanations that do not give counterfactual information on irrelevant features if, at stage (1), the explanatory context did not regard none of those features contextually salient.

Another point Rice's account diverges from Woodward's is the introduction of non-causal explanations alongside causal explanations, and unification of causal and non-causal explanations under the same account of explanation. Woodward's theory of counterfactuals is restricted to apply only to causal explanations (Woodward, 2003). But causal explanations are not so suitable for Rice's understanding of models. As we discussed in Chapter 2, it is generally not possible to represent difference-making causes accurately in models. But then, it is hard to claim that the counterfactual relationships produced by models are also causal in character. Accordingly, Rice (2021) distinguishes between causal claims (or what causes the explanandum) and explanation of the explanandum (or what explains the explanandum). In other words, for Rice, the counterfactual relations might not be necessarily causal, but still explanatory.

This distinction allows non-causal elements, like structural, mathematical, or statistical properties of the system, to produce counterfactual information and explanation. For example, consider the equilibrium explanations that are mainly associated with optimality models (Sober, 1983). Particularly, growth models that employ optimization techniques can demonstrate how the equilibrium point of per capita output counterfactually depends on borrowing constraints within the economy and consumption-saving trade-offs involved in consumer's budget spending. Such equilibrium explanation is non-causal as "where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired" (Sober, 1983, pp. 202). Put differently, growth models do not produce an explanation on how the resulted per capita output come to be, but it gives us a wide range of trajectories that lead to the same per capita output by combining different combinations of borrowing constraints and consumption-saving trade-offs.

The separation of "What is causal?" from "What is explanatory?" also helps him to keep the causal explanations under his account and maintain a pluralistic stance in his unified account. In rare cases where a model manages to justifiably employ the decompositional strategy, the explanation it produces would still be valid for Rice.

How about true counterfactual information? In order for an explanation that produces the necessary counterfactual information about dependence and independence to be successful, the counterfactual information should be true. *A* way to ensure the truth of the counterfactual information and the relevance of the model's result for the explanandum is universality. A universality result suggests that the model result matches with its target system, and hence with the explanandum, in its general patterns and overall behaviour. The next chapter elaborates on universality extensively.

3.1.2. Potential Criticism: Does this collapse into a decompositional strategy?

Going back to the earlier point about Rice seemingly adopting some version of the model decomposition assumption, there is indeed some tension there. Let me be more explicit. In the previous chapter, we saw that Rice's argument on why the role of relevant parts or difference-makers in the target systems cannot be translated into the model systems by means of accurate representation. But now, we point out that models can produce counterfactual information about the relationship between explanandum and contextually salient features of the target. For this, the model and the target need to share these contextually salient features. Such features can also be mathematical properties like symmetry, locality, homogeneity of certain parameters, not necessarily well-identified causes. Doesn't this reduce Rice's account to a decompositional account which suggests models explain due to their common features with their targets?

Similar versions of this criticism are also voiced against Batterman (2000) and Batterman and Rice (2014) by Lange (2015) and Reutlinger (2017). The debate on this still active (see, for instance, McKenna (2021) defending Batterman and Rice (2014)). My aim is not to get into the specifics of the discussion, but rather to point out this tension and pick up the hints Rice (2021) drops for a potential defence. The reason for this rather interpretative engagement is that Rice does not directly argue against the claim that his treatment of contextually salient features collapses into a decompositional account.

Initially, his account does not state that there is not going to be *any* shared features between the model and its target. There might be certain shared features, but Rice's holistic distortion view does not require them to be accurately represented in the model. More importantly, Rice separates difference-makers from features that are explanatorily relevant: "The explanatory context can make features that are not difference-makers explanatorily relevant" (Rice, 2021, pp. 103). Indeed, certain mathematical features can be shared across model and target systems. But Rice would discuss that they would not be sufficient for explanation. This is because any explanation based only on relating features across systems is not interested in explaining why all the other potentially relevant features (that are identified prior the modelling exercise) turned out to be irrelevant. In other words, such explanations are *not* constituted by a set of information about how the explanandum is counterfactually independent of these ex-post irrelevant features. So, what features they count as explanatory and how these features relate to the model and its target are different.

The thing that I struggle to understand is the following: In the previous chapter, we mentioned how Rice argues that it is really hard to replace or remove different (and idealized) parts and assumptions in a model without affecting the model result. Accordingly, Rice should acknowledge that both (a) (the ones that the target phenomenon counterfactually depends on) and (b) (the ones that the target phenomenon is counterfactually independent of) are required to get a model result in most cases. However, when the features in (b) are changed (possibly by assuming different parameter values) or removed, the model result stays roughly the same by definition of counterfactual independence. But this suggests that it is indeed possible to change these features in (b). Doesn't this make the necessity of features in (b) questionable?

This is different from the previous criticism reducing Rice's account into a decompositional account. And I do not think there is an incompatibility in the strict sense but more like an argumentative dilemma here: Arguing both that models' different features are not easily changeable without affecting the model result and that it is possible to get modal information about counterfactual independence seem difficult to reconcile simultaneously. If models are not decomposable and reversible, why do we care about the counterfactual independence in the first place? In my reading of Rice, I could not find a way to present and state his views in a way that overcomes this problem. So, while I could make a case for Rice to overcome the previous criticism, I do not see a way out for this.

If it is indeed possible to get information about counterfactual independence, the criticism of the decompositional strategies based on attacking D2 and D3 is weakened. This is because arguing that there are certain features that are counterfactually independent of the explanandum is not so different from saying that the contributions of the features in (a) to the result can be isolated from the contributions of the features in (b). These claims are not incompatible as a group because the features in (b) cannot be removed from the model to completely isolate the contributions of features in (a). However, I believe it weakens the strength of the criticism of decompositional strategies that we analysed in Chapter 2 to some degree.

3.2. Rice's Account of Understanding

After Rice's account on explanation, let us continue with understanding. His account of understanding has two necessary conditions: (i) grasping the systematic relationships between a fairly comprehensive set of information about the analysed phenomenon, and (ii) detecting how this information is related to diverse pieces of existing knowledge in the discipline (Rice, 2021, pp. 249). Here, grasping is used in a straightforward way as an activity of scientists in the sense of digestion or internalization of information. But (ii) helps attributing this activity not only to individuals, but also groups and communities of scientists as the existing knowledge could be communal: It is likely that communities have a larger set of background information about the studied phenomenon. Through different scientific activities like theorizing, modelling, experimentation, etc., new information about the phenomenon is extracted. Even though this new information is incorporated into the total body of information the scientific community has about the phenomenon, not every interconnection between this new piece of information and the existing body of information might be grasped by individual scientists. Still, each can contribute by providing "overlapping sets of information and connections that will constitute the scientific community's overall understanding of the phenomenon" (ibid: pp. 249).

Rice classifies his view of understanding as factive, meaning that some, but not necessarily all, of the information or beliefs within a body of understanding needs to be true (Rice, 2021). It is because he thinks "in order to genuinely understand a natural phenomenon *most* of what one believes about that phenomenon—especially about certain contextually salient propositions—must be true" (ibid: pp. 205). It naturally raises the question of "How much must be true?" As an answer, Rice recommends a "a case-by-case approach". So, how much of our beliefs about the phenomenon to be true for achieving understanding depends on the context. This is because it is hard to impose a contextually independent factive requirement for understanding salient features of the phenomenon that depend on the context: "in various contexts of scientific inquiry, some pieces of information or connections will be more salient, and so their truth-value or accuracy ought to carry more weight in determining whether one's understanding is factive" (ibid: pp. 252).

Such factivism is context-sensitive, quite flexible, and hence, considerably weaker than other factivist accounts. The aim with less strict factivist accounts is to show that idealized and distorted models can still be used to understand phenomena (Cornelissen & De Regt, 2022). Accordingly for Rice, who defends a holistically distorted view of models that can be detached from reality completely, the aim is to argue that non-realistic and hypothetical models can still produce factive understanding. That's why Rice defines factivism in a way that the truth requirement only applies to beliefs about the target, instead of models and theories directly. As we will see in Section 4.2.4 in the next chapter, this allows completely distorted and hypothetical models (e.g. Schelling's model of segregation) and false theories to still provide understanding. Of course, it is possible to argue that Rice faces a huge price for accommodating the contributions of idealizations and distortions understanding; that is, his account's relationship to truth is so weak that it is closer to non-factivism (for a similar argument, see Cornelissen and De Regt (2022)).

How about the content of these (mostly) true beliefs? Following explanation, Rice (2021) contends that the crucial kind of information in understanding is about counterfactual dependence and independence relations. So, "scientific understanding is constituted by a body of mostly true beliefs about how changes to the features of the system would (or would not) alter the phenomenon of interest" (Rice, 2021, pp. 256). Then, as explanations produce a large variety of such modal information, grasping an explanation enhances understanding significantly. Although this establishes a clear link between understanding and explanation, Rice (2021) maintains that explanation and understanding are not logically related.

In other words, explanation is neither sufficient nor necessary for understanding. It is not sufficient because it is possible to not fully grasp the explanation or to grasp it incorrectly. So, it is either due to the individual lacking the necessary background knowledge or simply his failure to interpret the counterfactual information captured in the explanation correctly. It is not necessary because even though a model fails to explain (e.g. because it is not able to provide all the counterfactual information about dependence and independence relationships), the *incomplete* explanation and the modal information can still contribute to understanding.

3.3. Conclusion

To make sense of Rice's pluralistic, flexible, context-specific account, we should remind ourselves what Rice takes as the main goal in offering a philosophical account of idealized modelling: "to justify scientists' use of these holistic distortions in terms of the explanations and understanding that they enable them to achieve that would otherwise be unobtainable" (Rice, 2021, pp. 284). Following this, we saw that his view of models as holistically distorted entities recognizes the limitations of modelling practice due to mathematical constraints and the necessity of distortions as tools to get model results in Chapter 2. Different real-world phenomena are likely to require different approaches to concretize and simplify the analysed entities and processes. So, the specific reasons for introducing the idealizations and the mathematical frameworks these idealizations induce are likely to be different across contexts. Then, there will be at least some variety of modelling techniques within and across disciplines.

This variety will inevitably have an impact on the ways the scientists use idealized models to explain. Subsequently, Rice (2018; 2021) stress that we need a pluralistic approach to how idealized models grant access to explanatory information. Some of them can be considered as causal explanations while others can be non-causal. Rice's counterfactual theory tries to unify both under a broadly factive view of explanation. It suggests that an explanation is successful when it provides true information about the analysed explanandum. The key is that the true information does not need to be about the explanandum's actual causes: In most of the cases, Rice claims that it will be about counterfactual dependence and independence of contextually salient features of the phenomenon. These counterfactual relations are not necessarily causal.

This also holds for understanding. To restate, "scientific understanding of a natural phenomenon is constituted by a community's or agent's grasp of a set of modal information about possible states of the system that is incorporated into a body of information that is mostly true" (Rice, 2021, pp. 298). Because of the special emphasis on information about counterfactual dependence and independence, explanation and understanding are strongly connected. But this connection does not translate into a logical connective as explanation is neither necessary nor sufficient for understanding. The connection, but not implication, between understanding and explanation emphasizes the various kinds of diverse routes with which scientific understanding can be achieved. Rice (2021) mentions investigating necessity claims, modelling hypothetical scenarios and exploring possibility space as an example of routes that produce understanding without (complete) explanation.

Having reconstructed what kind of learning (explanation and understanding) and information (modal information) models can provide about their targets, it is time to analyse Rice's argument on universality and how models can produce true counterfactual information about the phenomenon.

Chapter 4: Justifying Idealized Models with Universality

The previous chapter shows that idealized models need to produce counterfactual information about their target systems in order to be explanatory and provide understanding. We also saw that Rice's account of explanation is factive, meaning that any such counterfactual information must be true. Similarly, his account of understanding is also factive but more broadly compared to the explanation; *most* of the counterfactual information grasped by the individual or the community needs to be true. Now, the question is: How do idealized models provide true counterfactual information about their system? Put differently, where does their relevance and the irrelevance to the target phenomenon come from? Without giving a satisfactory answer to these questions, it is not possible to claim that the use of holistically distorted models is justified for the purposes of explanation and understanding. Rice's answer is built on the concept of universality. In this chapter, I will introduce and critically evaluate the concept of universality for the purposes of explanation and understanding from idealized models.

4.1. Universality: Basics

The switch in emphasis from models as separable entities, whose parts contribute to the result differently, to models as holistic distortions also implies a switch in emphasis from a detailed microscopic structure to general macroscopic behaviour. If the models are completely distorted representations, it is not possible to establish a connection between the microscopic structure of the target and the fundamental components, interactions and other features of the model for the purposes of learning. This also makes discovering any kind of causal relationship between the difference-making features of the target and the target result difficult, and possibly infeasible, *by analysing the model system*. To repeat what I argued before, without being able to represent difference-making causes accurately in models and to isolate their contributions to the result, it is hard to claim that there is any relevant causality for the target. The other option is to establish a connection between the model and the target via concentrating on the overall patterns and the general behaviour of the two systems. And universality is an instance of such a strategy.

As we are going to see in detail in Section 4.2.1, universality is a concept originated in theoretical physics (e.g. phase transitions) to describe the stability of certain features and parameter values (e.g. critical exponents) across different systems whose microscopic structures are completely different (e.g. fluids and magnets) (Morrison, 2015). Rice generalizes this concept and use the term to "simply mean the stability of certain patterns or behaviours across systems that are heterogeneous in their features" (Rice, 2021, pp. 155). Accordingly, universality classes are just the set of systems (model or target) which demonstrate those universal behaviours and patterns. The example of universality in phase transitions connect the behaviour of fluids and magnets, two diverse sets of systems, in the same universality class. Rice emphasizes that although the components, features, interactions and dynamics of those systems are distinct, the universality result ensures that they will demonstrate the same general patterns of behaviour. Put differently, even though different systems have different microscopic structures, their macroscopic behaviour will be similar within a universality class. This suggests that various microscopic structures might exhibit the same macroscopic behaviour.

Since Rice's main aim is to come up with *a particular* way to establish connections between the model and the real-world system, we can exploit this notion of universality: If we can show that the model system and the target system are in the same universality class, the counterfactual information coming from the model is ensured to be relevant for the target system as well (Rice, 2020). The argument goes as follows: Recall that it is implicitly assumed that scientist identifies contextually salient features of the explanandum before the modelling exercise. The explanandum and its contextually salient features are represented in the holistically distorted models that use idealizations as tools to extract model results. If there is a universality result, it means that the model result matches with the explanandum in its general patterns and overall behaviour. The initial set of contextually salient features, which have been presumably distorted within the model system, are now separable into (a) the ones that the target phenomenon counterfactually depends on and (b) the ones that the target phenomenon is counterfactually independent of.

The features in (a) are probably going to be non-causal elements like structural, mathematical, or statistical properties of the systems. So, the universality result potentially extracts information about what happens when there is a change in such properties or parameters. More importantly, the universality result reveals lots of modal information about why the features in (b) are irrelevant to the explanandum. This is crucial for universality-based explanation and understanding from models as it shows that most elements constituting the microscopic structures of different systems are not relevant for the overall behaviour. Thus, a particular universality result can grant explanation or understanding depending on how well it reveals counterfactual information about (a) and (b). A successful case justifies the usage of the holistically distorted system as they extract the stable and universal features of the whole class.

It is important to note the details on how Rice integrates universality to his account of models. Firstly, he does not claim that discovering universality classes is the only way to establish a relationship between the target and the model systems and his account allows pluralism in how we can learn from holistically distorted models (Rice, 2020). Secondly, he takes universality as a given empirical fact to link models with their target systems (Rice, 2019). He mentions that "appealing to the existence of a universality class in order to justify the use of an idealized model to explain is crucially different from providing an explanation of universality" (Rice, 2020, pp. 833). As such, he does not talk about how to explain universality classes that have been already discovered. I view this as a fair stance as it is not easy to develop an account of universality which both justifies the usage of holistically distorted models and explains why we observe these universality classes in the first place.

However, Rice (2019) also mentions that "scientists can justifiably use idealized models within a universality class to explain the behaviours of real-world systems in that class even when they fail to have a complete explanation of why that universality class occurs" (pp. 201). This gives universality classes an interesting status: Appealing to them is a way of connecting the target and the model systems and scientists can discover them "whenever two systems show an unexpected or deeply rooted identity of behaviour" (Kadanoff, 2013, pp. 178, cited as in Rice (2020)). But neither scientists nor philosophers have the burden of explaining why we observe this unexpected similarity among the macroscopic behaviours of two systems; it is an observation. I will discuss the problems with this aspect of universality extensively in Section 4.3.

4.2. Examples of Universality Results

To recap, universality occurs when different systems are noticeably similar (whatever it means) in their macrolevel patterns. Next, I will give two examples of universality results that Rice mentions. The first one is the aforementioned universality result in phase transitions that connect the behaviour of fluids and magnets. Analysing it is helpful to get an idea of the origins of universality. The second one is the only economics paper that explicitly studies universality: Parunak, Brueckner and Savit (2004). After presenting the paper and its result, in Section 4.2.3, I will critically discuss it in comparison with the case of universality in phase transitions. This will be helpful to illustrate the potential pitfalls of establishing and using the concept in a completely different domain than its origin in phase transitions. For the third one, I will analyse Schelling's checkerboard model of residential segregation, which is studied in Rice (2021) for the purposes of understanding, in Section 4.2.4. Examining these three examples will help me to identify more general problems and overall limitations of universality-based learning from models that I will discuss in Section 4.3.

4.2.1. Origins of Universality: Universality in Phase Transitions

Universality is a notion originated in theoretical physics. It has been especially discussed in the field of statistical mechanics, which employs idealized models often. As a technical term, it "describes the behaviour of the *critical exponents* associated with a *continuous phase transition*" (Parunak et al., 2004). A phase transition is a mathematical singularity that occurs in a system as some parameter changes (Morrison, 2015; Parunak et al., 2004). A singularity is essentially a point where the mathematical object at hand is not defined or well-behaved. In a phase transition, as a parameter varies, the system exhibits properties like discontinuity and non-differentiability. As a result, an infinitesimal change in a parameter leads to a qualitative change in the system. Consider the freezing of water where such a qualitative change occurs visibly; the change in temperature alters the density of water discontinuously at 1 atmosphere and 0 degrees Celsius. In this case, the density is called the order parameter.

Possibly the most popular example of a phase transition is the one of iron at the Curie temperature (760° Celsius) (Parunak et al., 2014; Morrison, 2015; Rice 2018). In this context, the qualitative change occurs in the net magnetization (i.e. the counterpart of density in the freezing of water) and it is the order parameter. Below the Curie temperature, iron is ferromagnetic, meaning that it is magnetized in the direction of the magnetic field to which iron is exposed and the magnetization remains even after the field is removed; above the Curie temperature, it is paramagnetic; the magnetization is proportional to the field and disappears when the field is zero (Parunak et al., 2004). The bottom line is that when iron is heated to a fixed critical temperature, it loses its magnetism, and hence the phase transition occurs.

The physicists analyse this transition from one state to another by modelling the path from ferromagnetism to paramagnetism as a function of temperature. Specifically, the behaviour of the order parameter near the critical temperature is described by a power function, where the power refers to the critical exponent: $|T - T_c|^{\alpha}$ (Morrison, 2015). Here, T is the temperature, T_c is the Curie temperature and α is the critical exponent. Interestingly, the value that α takes are relatively independent of the subject matter; Batterman (2000) mentions that the critical exponent for a lot of magnets like iron and neon is close to $1/3$ (cited as in Tieleman, 2022). What's more, the same functional form and the critical exponent is observed under other phase transitions as well, such as the vapour and fluid states of water (Tieleman, 2022; Elliott-Graves, 2022). This suggests that even though the underlying microscopic structure of magnets and water are diverse, they behave similarly close to the critical point. As a result, they are said to be in the same *universality class*.

For these diverse systems to behave similarly, the impact of their structures and the characteristics of their molecules on the macroscopic behaviour needs to be somehow neutralized. Why can this be the case? One of the main obstacles to modelling phase transitions is that the result of the observed transition (i.e. different states of water or iron) can only be achieved by introducing a key mathematical idealization to the model. Particularly, we know that the real systems that we aim to explain consist of finite number of particles (Morrison, 2015). However, the phase transition does not occur in the model world unless we assume that the system contains *infinitely many* particles (Morrison, 2015). The singular behaviour that we observe in the target phenomenon (e.g. the change in the density of water near 0° Celsius) requires an idealization on the number of particles. Once assumed, now it is possible to take *the thermodynamic limit* of the system as we approach to the critical point where the transition occurs. It allows the physicists to apply *renormalization* techniques and transform the Hamiltonian (or the total energy of the system). This procedure renormalizes the *irrelevant* features of the target system to zero, effectively neutralizing the characteristics of the molecules and their interactions (Parunak et al., 2004). So, the differences between, say, water and iron no longer affect the behaviour of the system around the critical points.

The universality result in phase transitions provides modal information about contextually salient features that the target phenomena counterfactually depend on via the power function. With the application of thermodynamic limit and renormalization techniques, it also provides information about why the behaviour of target phenomena does not dependent on certain features of the diverse phenomena (like structures and the characteristics of fluid and magnet molecules). As this is a universality result, it ensures relevance, and hence truth, of this set of counterfactual information about the set of explananda. As it satisfies all the conditions of Rice's account of explanation, this universality result, and its model characterized by the power function above, justifiably provides explanation.

This example concerned one of the first and most well-known universality results in physics and all other scientific disciplines. Other universality results Rice (2018, 2019, 2020, 2021) mention are from physics, biology and applied statistics. His successful cases of universality involve formation of melt ponds on sea ice sheets (Rice, 2018), Kardar-Parisi-Zhang universality class modelling biological growth (Rice, 2020), the universality of Gaussian distribution thanks to the central limit theorem (Rice, 2021). As one of my aims in this thesis is to analyse the applicability of Rice's framework to economics, let us analyse, to my knowledge, the only economics paper that explicitly studies universality and universality classes: Parunak, Brueckner and Savit (2004).

4.2.2. Universality in Economics: Multi-agent Systems

Parunak et al. (2004) work on multi-agent systems. A multi-agent system (MAS) is essentially a computational model for simulating the interactions of diverse agents to figure out the behaviour of the system under analysis. Their agents, which constitute the microscopic structure of the system, are not able to process every possible and/or available information in the best way. They generally have a certain number of predefined strategies and (as they face the same scenario/game many times) they adapt their preferences towards these strategies in time. However, this does not change the fact that they are still idealized models which are distorted in a way to utilize computational modelling techniques and get around the constraints related to computers and the programming language.[4]

The two systems they analyse are Minority Game and Graph Colouring. Minority Game refers to a model of repeated games in which an odd number of individuals choose one of the two decisions (say A or B) available to them each round. At each round, the individuals split into two groups: A-choosers, and B-choosers. The minority group wins the game. The decision of individuals in the minority group is awarded with a point (individuals receive no point if they are in the majority group). As there are multiple rounds, individuals need to devise strategies for what to do at every round of the game. To form their strategies, they are given some *m* past system states that describe the winner group before the game. It is possible to think *m* as the memory of the individuals. As *m* gets larger, more history the individuals can consider while making decisions. The individuals update their memory according to past information in the game and their preferences towards those strategies. Parunak et al. (2004) applies this setup to a resource allocation scenario where there are two suppliers, who can be overloaded if there is too much demand.

Graph colouring algorithms can also be used in such resource allocation problems. A graph is a mathematical structure which is made up of nodes that are connected by edges. Every node in the graph corresponds to a task, every colour symbolizes a resource and an edge of the graph correspond to the constraints between tasks. If there is an edge between two nodes and these nodes have the same colour, this resource [colour] cannot service the two tasks [nodes] simultaneously. Each node can take a colour from some finite set of colours and its colour can change over time. A node can perceive its neighbours' colours and any colour change is only perceived after some fixed time unit passes. At a given time with some given probability, the nodes can activate their local reasoning mechanism. When activated, a node re-evaluates its colour

[4] This is closely related to the notion of *computational tractability* introduced by Cherrier (2022).

assignment and computes the *Degree of Conflict (DoC)*. The aim is to minimize the proportion of adjacent nodes that have the same colour so that we allocate the resources [colours] efficiently. (Parunak et al., 2004, pp. 3-4).

These two systems are obviously different in their microscopic structure; agents' strategies, decision mechanisms and relationships with other agents are completely different across the systems. Now, it is important to note the variables that describe the macroscopic behaviour of the system over which a universality analysis can be made. In the Minority Game, Parunak et al. (2004) mention that the variance in the number of A-choosers (or B-choosers; the group does not matter) is a useful metric to analyse the system's general behaviour. As there is an inverse relationship between the total reward and this variance, this metric can be used to measure the system's efficiency.

For the graph colouring algorithm, the idea is to again measure the performance of the system. As such, Parunak et al. (2004) calculate the *global degree of conflict* (*GDOC*). This measure considers the constraint for a colour [resource] to be unable to service two nodes [tasks] simultaneously when these nodes are connected by an edge. Hence, they analyse the efficiency of the two systems in solving the resource allocation problems as the informational structure.

They state that these model systems manifest universality:

> Both models exhibit the same three regions: a region of low information with thrashing and herding, a region of excess information resulting in decisions no better than random, and a region where decision capability and available information are roughly balanced, yielding superior performance. These similarities arise even though the underlying decision mechanisms are very different (Parunak et al., 2004, pp. 4).

While this is not an example where diverse systems reduce to the same set of fixed points in their limit behaviour which can be expressed by a common functional form, the two systems still demonstrate the same set of behavioural regimes. The regime here corresponds to a specific set of configuration parameters. Particularly, these models show a similar pattern in their system efficiency as the configuration parameter values change. These parameters are the different values of *m* (the memory capacity of the individuals) in Minority Game and the different probability values of activating the local reasoning mechanisms of the nodes in the graph colouring algorithm.

Over the last two subsections, we have analysed two diverse universality results: one in phase transitions, the other in multi-agent systems. In the next subsection, I will critically discuss the universality result in Parunak et al. (2004) and argue that it is not suitable for the purposes of universality Rice had; justifying the usage of these idealized models for analysing real-world phenomena.

4.2.3. Discussion of Universality in MAS

To start with, it is important to note that Rice (2020) explicitly takes both the universality under phase transitions and MAS as genuine cases of universality and evaluates them in the context of justifying the usage of idealized models.

Still, bracketing Parunak et al. (2004) with examples in physics (e.g. phase transitions) and biology (e.g. Kardar-Parisi-Zhang universality class analysing biological growth) might not be straightforward. Initially, let us zoom in on why Parunak et al. (2004) find universality important. They briefly talk about the importance and potential of universality in constructing models of the real world and particularly agent-based models (pp. 6-7). However, their main discussion centres around the benefits of discovering universality classes on the system design and implementation. If we have two models with individuals whose decision mechanisms are simple and complex respectively, and individual interaction leads to the same overall results for the resource allocation problems, it is possible to simplify the model design. Such a simplification also reduces the engineering costs significantly by removing the sophistications that do not have an impact on the results (Parunak et al., 2004).

So, the focus is not directly on justifying the usage of these idealized models for analysing real-world phenomena. In fact, it is not clear to what extent the behavioural patterns generated by the Minority Game and the graph colouring algorithm are tied to the decision-mechanisms that are observed in the real-world. The qualitative changes that occur under the phase transitions of water (i.e. the change in density) and iron (in the state of magnetization) are facts about the real-world phenomena. Within statistical mechanics, universality is a property of model systems whose relationships to their respective target systems are well-documented and tested with extensive empirical studies.

This is not the case for the universality between the Minority Game and the graph colouring algorithm. Here, we essentially describe a universality case between two models whose relationships to the real-world phenomena have not been properly established. As we have seen in Chapter 1, such relationships are mainly achieved by formulating a target system that specifies some parts of the real world. And there is not any explicitly formulated corresponding target system for these MAS models to represent. Then, even though these two models might be in the same universality class, it is not clear why this is enough to also conclude this instance of universality establishes a connection between the idealized model and the real-world. This might not be even part of the aims of the authors as the study is not motivated by empirical concerns but computational concerns like system design and efficiency.

To make my point clearer, compare this with the universality under phase transitions. The former is essentially a robustness result which states that a change in the decision mechanisms does not change the outcome across different *model systems* since it is neutralized by the interactions between the individuals and their ability to process information. The latter concludes that certain macroscopic behaviours (i.e. qualitative changes in molecular and atomic properties near specific points) in two diverse target systems can be captured by the same functional form and are reducible to the same set of fixed points under their respective model systems. Thus, the two target systems and the model system describing their macroscopic behaviour are in the same universality class. These are distinct results with different relationships to the world.

It is important to clarify that I do not claim that the result in Parunak et al. (2004) is useless or worthless. It has value for the purposes of system design, and it might pave the way for other future models that produce explanation and understanding. Given this, it is indeed possible to claim that this justifies the idealized models of MAS. However, whatever we learn about the system design is not related to Rice's two main ways of learning, which are explanation and understanding. The reason is that, again, this universality class does not cluster Minority Game and Graph Colouring with a target system. So, the produced counterfactual information about dependence and independence of features cannot be related to an actual phenomenon. Therefore, I do not agree with Rice on the significance of the universality result in Parunak et al. (2004), even though it might be beneficial for computational concerns regarding MAS.

4.2.4. Schelling's Checkerboard Model of Segregation

Next, let us focus on how holistically distorted models can provide understanding without providing explanation. Rice (2021) gives Schelling's checkerboard model of segregation as an example of such models. While talking about segregation of two types of individuals in cities (say A and B), one possible and convincing mechanism that could lead to it is strong discriminatory (or racist) preferences. In Schelling's model (Schelling, 1969), nickels and dimes are taken to represent A and B types respectively. An agent's neighbourhood is taken to be a checkerboard of some size, such as a 3 x 3 board with a set of nine adjacent squares. The model assumes some mild form of discriminatory preferences in that any individual wants at least 30% of their neighbours to be in the same type as them. Each agent then takes turns to remain in the same square or move to another depending on whether their preferences are satisfied or not. So, an agent can either remain where he is or can move to an unoccupied adjacent square such that his new set of neighbours satisfy his preferences (i.e. wanting 30% of your neighbours to be type A if the person is type A). Given some random allocation of agents over the board, this continue until all agents are satisfied with their locations.

It is not so easy to relate this model to an actual city, or a real-world target system as it is extremely simplified and highly distorted. Setting aside the problems with the target, Schelling's model shows that segregation can arise even if the individuals have only mild discriminatory preferences (i.e. wanting 30% of their neighbours to be in the same race as them). Besides, the result holds across a wide range of robustness checks in individual's preferences, the checkerboard size, spatial configurations, and the rules for moving to another location. So, Rice (2021) interprets this as an *unnecessity result* (or a possibility result); unlike the common belief at the time, strong discriminatory preferences are not necessary for residential segregation to occur.[5] As Schelling also mentions, his aim was not to represent any particular city but to investigate "some of the individual incentives and individual perceptions of difference that can lead collectively to segregation" (Schelling 1978, 138). Thus, his holistically distorted model shows that segregation is possible even if everyone has some small preference towards homogenous neighbourhoods (Rice, 2021).

Can this model provide explanation? If not, can it provide understanding? Rice's answer to the first question is negative but to the second is affirmative:

> Merely knowing that it is possible for mild preferences for like neighbours to produce segregation falls short of being able to explain why cities are actually segregated. However, by showing that a particular set of preferences could possibly give rise to the phenomenon, Schelling's model is able to produce some information that enables us to better understand the phenomenon of interest. Specifically, this model is able to justify the true belief that a neighbourhood can become segregated even if there are no strong racist preferences (i.e., individuals acting on strong racist preferences is *not* a necessary condition for segregation to occur). This information is enlightening even if this fact is not part of the actual explanation of why cities are segregated. Therefore, Shelling's checkerboard model produces some understanding by undermining a formerly accepted claim about what was necessary for segregation to occur, but it does not even aim to provide an explanation of that phenomenon (Rice, 2021, pp. 236-237).

The results of Schelling might justify the belief in the counterfactual claim that, all the other factors remained the same, even without strong discriminatory preferences, segregation would still occur (in some abstract context). Clearly, this modal information produced by the model does not reveal any feature that a possible target phenomenon (i.e. actual cities) counterfactually depends on. And it only cites *one* counterfactually irrelevant factor. Grasping this is not sufficient for "a complete

[5] In some sense, this is the opposite case of theoretical *impossibility results* (e.g. Arrow's impossibility theorem) showing that a particular set of problem(s) cannot be solved as stated in the claim.

explanation of how any actual segregated city has arisen" (ibid: 236). But it is sufficient for understanding.

The problem is that it is not so clear what else would be necessary to explain residential segregation in an actual city. Are we lacking necessary modal information? Is explanation always unattainable from this model due to Schelling's more modest aims? Or is it unattainable because it is so abstract that it does not have an actual target? I interpret Rice's stance in the quote above as signifying the idea that it does not provide explanation because it lacks modal information.

This leads to a question for which I could not find any answer when reading Rice. It is about the relationship between explanatory completeness and usefulness. Given that pragmatic considerations play a really large role in Rice's account of explanation, it is not clear to me why an explanation first needs to be complete to be counted as useful. I do not claim that Schelling's model is explanatory; I do not think it is. But I just do not understand while some set of information cannot constitute an explanation that is partial or incomplete but useful, it can constitute a perfectly fine understanding. That's why I believe that it is better to focus on the fact that model being abstract and targetless and that hinders in the explanatory potential of Schelling's model.

Concluding this section, I gave examples of three models that Rice explicitly refers to for purposes of universality-based explanation and understanding without explanation. In the next section, I will expand on the drawbacks associated with Rice's usage of universality in justifying the use of holistically distorted models.

4.3. Drawbacks of Universality and Universality-Based Learning

In this final section, I will discuss some of the general drawbacks and potential limitations of Rice's treatment of universality. My aim is to develop further the points I was hinting at towards the end of section 4.1 and to draw some general lessons from the previous section. Before we start, let us remember the definition of Rice: Universality is "a statement of the fact that different physical systems will nonetheless display similar patterns of behaviour that are largely independent of their physical features" (Rice, 2020, pp. 832). Accordingly, Rice takes universality as a given empirical fact and whenever it is discovered that a model system and its corresponding target system are in the same universality class, the model can be used to explain the target even without an explanation of why that universality class occurs.

Broad Definition. The main problem with this expanded version of universality is that it is too broad and does not specify any methodological requirements to identify universality classes. In its original domain, the similarity within the universality class is measured in terms of atomic and molecular properties and among system behaviours in terms of critical exponents. Compare this with Parunak et al. (2004) where the variable of interest is system efficiency. It is different than the order parameters under phase transitions (the net magnetization of iron or the density of water) in that it is more suitable for analysis within model systems, rather than establishing a relationship with the world via a target system. The reason is that it is quite complex to define and measure a counterpart of the system efficiency for real-world systems of decision-making.[6] As the concept of universality generally arises in *some* set of parameters *near* certain points, like the limit cases, it is not clear how we can systematically identify the parameters and regimes on which we analyse the similarities of behavioural patterns across different systems.

This absence of a clear-cut methodology or guiding principles to identify universality classes surely increase the applicability of universality in different contexts. The trade-off is that the set of possible universality results enlarges to a point where it is highly likely that they will be connected to their real-world targets in varying degrees (compare universality in phase transitions and universality in MAS).[7] Hence, universality-based learning from models can be quite limited and uninteresting for the purposes of model explanation and understanding.

Models without Targets. In Section 4.2.3, I argued that MAS systems in Parunak et al. (2004) do not have a respective target system, and this limits the explanatory power of the universality result. This case can be generalized: Models without explicitly formulated target systems can easily lead to the discovery of universality classes whose models do not provide explanation and understanding, understood in the sense of Chapter 3. Such concerns about targets are highly related to the general-vs-generic use of models, introduced by Elliott-Graves (2022).

General models are meant to be applicable to many target systems. Having a large set of targets to which the model applies is valuable because it helps to analyse seemingly distinct phenomena under the same model: "Showing a particular phenomenon is an instance of a more general pattern because it is caused by a common set of rules or mechanisms can help scientists explain the phenomenon or predict how the system will behave in the future" (Elliott-Graves, 2022, pp.4). This can be related to the goal of Rice in establishing universality classes: If we have two distinct phenomena whose corresponding target systems are in the same universality class with a unique

[6] This lack of quantitative metrics to study universality in MAS is also emphasized by Parunak et al. (2004, pp. 2).

[7] To give credit, Parunak et al. (2004) also speculate about a possible hierarchy between universality results and rank their result lower than the one under phase transitions.

model, the model can be used to transfer knowledge (i.e. modal information) from one target system to another.

On the other hand, generic models are not meant to be applicable to *any* real-world system. It is used to capture hypothetical and abstract models that are highly distorted. Schelling's segregation model that we analysed in Section 4.2.4 is a good example of generic models. Recall that Schelling's aim was not to represent any specific city with this model (Aydınonat, 2007), but to uncover an abstract mechanism for segregation which does not provide learning about real-world residential segregation (Fumagalli, 2016, cited as in Elliott-Graves (2022)).

This distinction is relevant because we should not appeal to universality-based explanation for models that are used generically. Without a target system, a universality result does not reveal anything relevant to the real-world phenomena. This point is important because it might be tempting to interpret generic models as general models and claim universality results. By combining other generic models with the broad definition of universality, it is possible to have ill-classified universality classes and overgeneralized models.

This potential problem might be exacerbated since some systems lend themselves to both general and generic model use (Elliott-Graves, 2022) and Rice's universality account cannot distinguish them. In fact, the Ising model which describes the phase transitions in Section 4.2.1 can be used in both ways: While it can illustrate the abstract phenomenon of phase transition (with no reference to specific microscopic details of any target system), it has been also revealed that it applies to many real-world systems like freezing of water or magnetism of iron (Elliott-Graves, 2022). This example shows that a generic model can be used as a general model. However, this does not mean that generic models can be used as general models frequently. The problem is that scientists are not always explicit about the intended uses of their models, how their models relate to an actual phenomenon and how the model results should be interpreted. Then, one should be cautious in generic-vs-general distinction when attributing universality results. Otherwise, it is possible to wrongly establish universality classes among some generic model and seemingly related real-world phenomena, and to use this universality result for the purposes of explanation and understanding.

Unexplained Universality Results. Next, I will expand upon the problems of appealing to universality without an explanation on why that instance of universality occurs. I argue that such use of universality can again produce ill-classified and arbitrary universality classes. The issue stems from the fact that it is not straightforward to distinguish whether the observed common macroscale behaviour is generated by the systems in the universality class, or it is just a coincidental similarity coupled with empirical content.

To demonstrate it, I believe it is useful to compare universality with a highly related concept called *universal patterns*. A pattern is a structure (as it is perceived as being structured) with no empirical content; it cannot be empirically true or false (Tieleman, 2022). Think about geometric shapes, fractals or series that can be expressed with a functional form; such patterns can be related to facts about phenomena when they are coupled with empirical content.

For instance, the normal distribution can be used as the distribution of height. The observed height levels are the information that give meaning to the symmetrical bell shape pattern of the normal distribution. However, it is also possible to couple the symmetrical bell shape pattern with other empirical content, such as the weight of loaves of bread (Lyon, 2014; cited as in Tieleman (2022)). Then, a pattern is said to be a universal pattern if and only if "it can be made to refer to facts about phenomena in multiple domains by changing just the empirical content that the pattern is coupled with" (Tieleman, 2022, pp. 9).

What is the difference between universal patterns and universality? After all, we can think about the power function $|T - T_c|^{1/3}$ as a universal pattern that can be made to refer to facts about state transitions in water and transitions in magnetism of iron. The difference is that under universality, the system behaviour comes in the form of the observed pattern (Tieleman, 2022). As such, the system (or the class of systems) *generates* the observed pattern, or the empirical content. Yet, the notion of universal patterns does not necessarily tie the observation to the system it is generated by; there is nothing specific about the systems of human height and the weight of loaves of bread so that we observe normal distribution.

This also underlines the differences in purposes. The universal patterns are discussed in the context of *model transfer* and the important thing is to check to what extent the data of different phenomena fit into the universal pattern. This concerns the usefulness of the pattern for goodness of fit. However, Rice discusses universality in the context of *model explanation and understanding*. Combining the broad definition of universality with no burden of explaining why the systems in the same universality class generate the observed pattern might lead to overgeneralization and arbitrariness. It is not clear why we should not perceive an instance of unexplained universality any different from the similarity between human height levels and the weight of loaves of bread. Under both cases, it is not clear why some piece of counterfactual information about a phenomenon (human height levels) will also be relevant for a completely different phenomenon (weight of loaves of bread) under the same universality class. Without anything that tie the similarities in general behaviour of different phenomena, I do not see what separates an unexpected universality result from a coincidence. To avoid this, someone, either philosophers or scientists, should be given the task of elaborating why a particular universality class occurs. Otherwise, Rice's account of universality fails to be any different from universal patterns.

That said, I don't want to argue for a requirement of complete explanation for universality results. A well-informed discussion on the potential drivers of the universality result would be sufficient. For instance, *a possible* explanation of the universality result we analysed for MAS in Section 4.2.2 is that the individuals are boundedly rational in both models. So, their ability to respond to information is limited and any sophistication of information above their limit just overwhelms their decision mechanism, effectively introducing noise to the system which reduces restricts system efficiency (Parunak et al., 2004). Then, just like the renormalization techniques neutralize the characteristics of the molecules and the impact of their interactions under phase transitions (Morrison, 2015), it is possible to hypothesize that rationality characteristics of the individuals limit the information that can be processed, and this neutralizes the effect of additional information. This also restricts the interactions that could occur among the individuals at different information levels. As a result, these diverse systems exhibit similar macroscopic behaviour.

Again, why we observe universality is another question that requires a separate analysis. But it seems like Rice, together with Robert Batterman, demanded more from a universality-based learning in an earlier paper: Batterman and Rice (2014). After analysing examples from fluid dynamics (Lattice Gas Automaton) and biology (Fisher's sex ratio model), they conclude that the models are explanatory because they have a *backstory* (i.e. the renormalization group strategy) showing that the model and its target belong to the same universality class. So, the universality result is explanatory only if (i) it is a discovered empirical result among the two systems, and (ii) there is a backstory which ensures that the model system will reproduce the macroscopic behaviour of the target system (Batterman & Rice, 2014, pp. 364). In Rice (2018; 2019; 2020; 2021), Rice drops condition (ii), and deems discovering a universality result among the model and the target sufficient to learn from the model (though it might not be sufficient for a complete explanation of the target). I do not understand why he loosens the requirement from Batterman and Rice (2014), but, as I argued, I find it problematic.

4.4. Conclusion

In this chapter, I introduced the concept of universality and demonstrated some drawbacks and limitations of it for model-based learning, especially for explanation. Rice uses a generalized version of the concept, compared to its origins in phase transitions, to mean "stability of certain patterns or behaviours across systems that are heterogeneous in their features" (Rice, 2021, pp.

155). Accordingly, universality classes are just the set of systems, model, or target, which demonstrate those universal behaviours in their overall, macroscopic patterns. Rice exploits this notion of universality to establish the relevance of the counterfactual information produced by the model system, and possibly use the same counterfactual information across different target systems in the same universality class. As a result, such counterfactual information can now provide both explanation and understanding about the explanandum, and hence, justify the use of holistically distorted models.

I mentioned two examples of universality results, one of phase transitions in physics (Section 4.2.1) and the other of multi-agent systems in economics (Section 4.2.2). These are helpful to illustrate the concept and how it transpires in different contexts. However, comparing them shows that they differ in their relationship with the real world (Section 4.2.3). Under the universality class in multi-agent systems of Parunak et al. (2004), it is not clear to what extent the behavioural patterns generated by the systems are tied to the actual decision-mechanisms. This is because there is no corresponding target system which connects the results of the model to real-world. As I explained, it might not be even part of the aims of the authors as the study is not motivated by computational concerns only.

Another example of a targetless model is Schelling's checkerboard model of segregation (Schelling, 1969), which is used generically and is not meant to be applicable to *any* real-world system (Section 4.2.4). I analysed Rice's argument on how this model still produces counterfactual information that is insufficient for the purposes of explanation, but sufficient for purposes of understanding. Based on this, I briefly talked about the unclarity in hierarchy between explanatory completeness and explanatory usefulness within Rice's framework.

Finally, generalizing on these illustrative examples, I pointed out the general drawbacks and potential limitations of universality. Particularly, appealing to universality is problematic when the model(s) at hand (i) is targetless (like the ones in Parunak et al. (2004)), (ii) it is meant to be used generically (Schelling's segregation model). In such cases, it is possible to have overgeneralized and arbitrary universality classes which do not serve the purposes of justifying holistically distorted models. Similar unwanted and misleading results can also be multiplied by (a) his definition being too broad without a clear methodology for identification of universality classes, (b) his account allowing the usage of universality justifiably without an explanation on why it occurs.

On a last additional note on the applicability of universality-based learning to economic models, I think its use would be limited. The prevalence of hypothetical and highly abstract models in economic theory makes it susceptible to the issues of overgeneralization and arbitrariness.

Moreover, even if the concept of universality was perfect without any problems, it would receive limited appeal in the fields of philosophy of economics and economic methodology. The reason is that it is not clear how philosophers of science or methodologists can make use of universality classes as a way of providing explanation and understanding from models when the concept is not discussed by the scientists explicitly. While such discussions exist in disciplines like physics and biology, in economic literature, the explicit mentions to universality and universal behaviour are rare (I could only find Parunak et al. (2004)). This makes using the universality argument of Rice to establish a connection between an economic model and an economic phenomenon difficult.

Conclusion

In this thesis, I reconstructed and critically analysed Rice's framework to answer the question of "How can we learn from a model about real-world phenomena when the model clearly idealizes the phenomena it is supposed to represent?" To do so, I first introduced target systems and idealizations in Chapter 1, focusing on how they are understood and used in the literature. My discussion points out two main things: (i) target system specification is not an area studied as commonly as model systems, and it functions as kind of a black box between the real-world phenomenon and the model, (ii) defining idealizations as distortions allows subsuming both abstractions and approximations into idealizations under a broad interpretation of what counts as a distortion.

Next, in Chapter 2, I started analysing the first pillar in Rice's framework: his criticism of decompositional strategies and holistic distortion view of idealized models as an alternative. I reconstructed, explained, and illustrated Rice's criticism of the philosophical accounts that employ the decompositional strategy. Based on the main conclusions from the criticism, I then explained Rice's alternative way of understanding models that characterize models as holistic distortions of their target systems. This view aims to justify the use of idealized models by emphasizing their ineliminable contributions to extracting model results, and hence to explanations and understanding. I also substantiated Rice's concerns for mathematical modelling by appealing to issues of integration and tractability in economic methodology. While this chapter is not argumentative in spirit, it underlines that Rice's holistic distortion view of models can be promising for economic models. I demonstrated its suitability to practice of model building in economics in general by appealing to various examples from different subfields of economics, including growth theory and game theory.

Then, I continued with the analysis of Rice's counterfactual account of explanation and understanding in Chapter 3. As the account is intertwined with many case studies and other theories of explanation, it is hard to see how this account is supposed to work in general, and how it fits into Rice's general framework of scientific models. To clarify these issues, I presented my own reconstruction of his account of explanation in a modelling scene step-by-step and expanded on the unclear and crucial concepts on each step. My reconstruction also helps to illustrate and discuss a criticism of Rice's account collapsing into a decompositional strategy and the argumentative tension in his framework due to conflicting aims of arguing for both that models' different features are not easily changeable without affecting the model result and that it is possible to get modal information about counterfactual independence. Finally, I gave the highlights of his account of understanding and how it differs from explanation.

In Chapter 4, I introduced the concept of universality and showed some drawbacks and limitations of it for model-based learning. I first explained how Rice uses the concept to show a particular way of models to provide true counterfactual information about the phenomenon. Then, I gave examples of three models that Rice explicitly refers to for purposes of universality-based explanation and understanding without explanation. While doing so, I also compared a two set of universality results: the one in phase transitions and the one in multi-agent systems. I argued that these do not constitute the same type of universality result due to the differences in their model systems' relationships with their corresponding target systems. Next, by generalizing on these three examples, I indicated that appealing to universality is problematic when the model(s) at hand is targetless, and it is meant to be used generically. In such cases, it is possible to have overgeneralized and arbitrary universality classes which do not serve the purposes of justifying holistically distorted models. This restricts the potential of universality-based learning in economics, given the prevalence of hypothetical and highly abstract models in economic theory. Finally, I argued that similar results might also arise because (a) Rice's definition is too broad without a clear methodology for identification of universality classes, and (b) his account allows the usage of universality justifiably without an explanation on why it occurs.

Lastly, a few words on what this thesis suggests about the potential venues of research are in order. Throughout the thesis, target systems and the unclarity of their nature and specification process led to problems in the analysis of idealizations, Rice's account of explanation and universality-based learning. This is an unexpected result, given that it is such a fundamental concept that most research in philosophy of science builds upon. Accordingly, I believe that this thesis shows both the necessity and importance of having a more comprehensive and integrated understanding of target systems to models and real-world phenomena. The vagueness of the current state of the notion, excepting rare cases such as Elliott-Graves (2020), hinders the progress of understanding scientific models. So, instead of moving forward with developing different accounts of learning to justify the scientists' use of idealized models, maybe it is better to take a step back and strengthen the base upon which these accounts are built.

References

Akerlof, G. A. (1970). The Market for 'Lemons': Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Aydinonat, N. E. (2007). Models, conjectures and exploration: An analysis of Schelling's checkerboard model of residential segregation. *Journal of Economic Methodology, 14*(4), 429‑454.

Aydinonat, N. E. (n.d.). The Puzzle of Model-Based Explanation. In T. Knuuttila, N. Carrillo, & R. Koskinen (Eds.), *The Routledge Handbook of Philosophy of Scientific Modeling*. London: Routledge.

Batterman, R. W. (2000). Multiple Realizability and Universality. *British Journal for the Philosophy of Science*, *51*(1), 115–145. https://doi.org/10.1093/bjps/51.1.115

Batterman, R. W. (2009). Idealization and Modeling. *Synthese*, *169*(3), 427–446. https://doi.org/10.1007/s11229-008-9436-1

Batterman, R. W., & Rice, C. C. (2014). Minimal Model Explanations. *Philosophy of Science*, *81*(3), 349–376. https://doi.org/10.1086/676677

Boumans, M. J. (1999). *Built-In Justification* (SSRN Scholarly Paper No. 1434348). https://papers.ssrn.com/abstract=1434348

Cherrier, B. (August 2022). The Cost of Virtue: Some Hypotheses on How Tractability Shaped Economics. Available at SSRN: https://ssrn.com/abstract=3927806 or http://dx.doi.org/10.2139/ssrn.3927806

Cornelissen, M. D., & Regt, H. W. de. (2022). Understanding in Synthetic Chemistry: The Case of Periplanone B. *Synthese*, *200*(6), 1–31. https://doi.org/10.1007/s11229-022-03929-y

de la Croix, D., & Michel, P. (2002). *A Theory of Economic Growth* [Cambridge Books]. Cambridge University Press. https://econpapers.repec.org/bookchap/cupcbooks/9780521001151.htm

Elliott-Graves, A. (2020). What is a Target System? *Biology and Philosophy*, *35*(2), 1–22. https://doi.org/10.1007/s10539-020-09745-3

Elliott-Graves, A. (2022). What Are General Models About? *European Journal for Philosophy of Science*, *12*(4), 1–26. https://doi.org/10.1007/s13194-022-00502-9

Friedman, M. (1953). The Methodology of Positive Economics. In M. Friedman (Ed.), *Essays in Positive Economics* (pp. 3–43). University of Chicago Press.

Fumagalli, R. (2016). Why We Cannot Learn From Minimal Models. *Erkenntnis*, *81*(3), 433–455. https://doi.org/10.1007/s10670-015-9749-7

Godfrey-Smith, P. (2009). Abstractions, Idealizations, and Evolutionary Biology. In A. Barberousse, M. Morange, & T. Pradeu (Eds.), *Mapping the Future of Biology: Evolving Concepts and Theories* (pp. 47–56). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9636-5_4

Hands, D. W. (2016). Derivational Robustness, Credible Substitute Systems and Mathematical Economic Models: The Case of Stability Analysis in Walrasian General Equilibrium Theory. *European Journal for Philosophy of Science*, *6*(1), 31–53. https://doi.org/10.1007/s13194-015-0130-0

Hindriks, F. A. (2006). Tractability assumptions and the Musgrave–Mäki typology. *Journal of Economic Methodology*, *13*(4), 401–423. https://doi.org/10.1080/13501780601048733

Jebeile, J. (2017). Idealizations in Empirical Modeling. In M. Carrier & J. Lenhard (Eds.), *Mathematics as a Tool. Tracing New Roles of Mathematics in the Sciences*. Springer Verlag.

Jones, M. R. (2005). Idealization and Abstraction: A Framework. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *86*(1), 173–218.

Kadanoff, L. (2013). Theories of Matter: Infinities and Renormalization. In R. Batterman (Ed.), *The Oxford Handbook of Philosophy of Physics* (p. 141). Oup Usa.

Knuuttila, T., & Morgan, M. S. (2019). Deidealization: No Easy Reversals. *Philosophy of Science*, *86*(4), 641–661. https://doi.org/10.1086/704975

Kuorikoski, J., Lehtinen, A., & Marchionni, C. (2010). Economic Modelling as Robustness Analysis. *The British Journal for the Philosophy of Science*, *61*(3), 541–567.

Lange, M. (2015). On ?Minimal Model Explanations?: A Reply to Batterman and Rice. *Philosophy of Science*, *82*(2), 292–305. https://doi.org/10.1086/680488

Levy, A. (2018). Idealization and Abstraction: Refining the Distinction. *Synthese*, *198*(Suppl 24), 5855–5872. https://doi.org/10.1007/s11229-018-1721-z

Lyon, A. (2014). Why Are Normal Distributions Normal? *British Journal for the Philosophy of Science*, *65*(3), 621–649. https://doi.org/10.1093/bjps/axs046

Mäki, U. (1992). On the Method of Isolation in Economics. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *26*, 19–54.

Mäki, U. (2011). Models and the Locus of Their Truth. *Synthese*, *180*(1), 47–63. https://doi.org/10.1007/s11229-009-9566-0

Mäki, U. (2020). Puzzled by Idealizations and Understanding Their Functions. *Philosophy of the Social Sciences*, *50*(3), 215–237. https://doi.org/10.1177/0048393120917637

McKenna, T. (2021). Lange on Minimal Model Explanations: A Defense of Batterman and Rice. *Philosophy of Science*, *88*(4), 731–741. https://doi.org/10.1086/713890

Morrison, M. (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. New York, US: Oup Usa.

Parunak, H. V. D., Brueckner, S., and Savit, R. 2004. Universality in multi-agent systems. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, ed. N. R. Jennings, 930–937. New York: Association for Computing Machinery.

Portides, D. (2013). Idealization in Economics Modeling. In H. Andersen, D. Dieks, W. González, T. Uebel, & G. Wheeler (Eds.), *New Challenges to Philosophy of Science* (pp. 253–263). Springer Verlag.

Portides, D. (2021). Idealization and abstraction in scientific modeling. *Synthese*, *198*(24), 5873–5895. https://doi.org/10.1007/s11229-018-01919-7

Portides, D. P. (2007). The Relation between Idealisation and Approximation in Scientific Model Construction. *Science & Education*, *16*(7), 699–724. https://doi.org/10.1007/s11191-006-9001-6

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Reutlinger, A. (2017). Do Renormalization Group Explanations Conform to the Commonality Strategy? *Journal for General Philosophy of Science / Zeitschrift Für Allgemeine Wissenschaftstheorie*, *48*(1), 143–150. https://doi.org/10.1007/s10838-016-9339-7

Rice, C. (2018). Idealized Models, Holistic Distortions, and Universality. *Synthese*, *195*(6), 2795–2819. https://doi.org/10.1007/s11229-017-1357-4

Rice, C. (2019). Models Don?T Decompose That Way: A Holistic View of Idealized Models. *British Journal for the Philosophy of Science*, *70*(1), 179–208. https://doi.org/10.1093/bjps/axx045

Rice, C. (2020). Universality and Modeling Limiting Behaviors. *Philosophy of Science*, *87*(5), 829–840. https://doi.org/10.1086/710623

Rice, C. (2021). *Leveraging Distortions: Explanation, Idealization, and Universality in Science*. The MIT Press. https://doi.org/10.7551/mitpress/13784.001.0001

Rubinstein, A. (1980). *Perfect Equilibrium in a Bargaining Model (Now published in Econometrica, vol.50, (1982), pp. 97-100.)* [STICERD - Theoretical Economics Paper Series]. Suntory and Toyota International Centres for Economics and Related Disciplines, LSE. https://econpapers.repec.org/scripts/a/abstract.pf?h=RePEc:cep:stitep:13;terms=rubinstein%201982

Schelling, T. (1969). Models of Segregation. *American Economic Review*, *59*(2), 488–493.

Sober, E. (1983). Equilibrium Explanation. *Philosophical Studies*, *43*(2), 201–210. https://doi.org/10.1007/bf00372383

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press. https://doi.org/10.2307/j.ctv1dv0tnw

Suárez, M. (2010). Scientific Representation. *Philosophy Compass*, *5*(1), 91–101. https://doi.org/10.1111/j.1747-9991.2009.00261.x

Tieleman, S. (2022). Model Transfer and Universal Patterns: Lessons From the Yule Process. *Synthese*, *200*(4), 1–20. https://doi.org/10.1007/s11229-022-03737-4

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. New York, US: Oxford University Press.

Woodward, J. F. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Zach, M. (2022). Revisiting Abstraction and Idealization: How Not to Criticize Mechanistic Explanation in Molecular Biology. *European Journal for Philosophy of Science*, *12*(1), 1–20. https://doi.org/10.1007/s13194-022-00453-1