Soundness Analysis: Justifying Deductive Reasoning in Model-Based Sciences

Raja Panjwani*and Aki Lehtinen†

April 27, 2023

Abstract

Scientists often draw deductive inferences from highly idealized models. These models characterize true features of their targets, but they also invariably misrepresent some of their attributes, often with explicit falsehoods. This article explains why this prima facie unsound practice can systematically lead to insights. We develop a method, which we call 'Soundness Analysis', for extracting from a model a set of true statements about a target system, which can be used to construct a sound argument for a model's conclusion. We illustrate the method on the Hoff-Stiglitz model from economics, and we argue that the possibility of such an analysis sheds new light on mathematical explanation in model-based sciences.

1 Introduction

An important virtue of model-based reasoning is its propensity to offer surprising insights about target systems of interest (e.g., Knuuttila 2009). In mathematical modeling, the results are typically expressed as theorems. However, the assumptions used in their derivations invariably make false claims about their targets: in addition to abstracting away irrelevant details, they idealize by distorting the properties in the target. The question thus arises: why give epistemic weight to theorems when the assumptions on which they

*New York University Stern School of Business, rp2662@nyu.edu.
†Nankai University, aki.lehtinen@helsinki.fi

are based are known to be false? If treated as arguments for a conclusion of interest, the proofs in question may be deductively valid, but they are also clearly unsound (e.g., Cartwright 2007). This practice is ubiquitous in economics, and also common in some parts of biology, physics, and other model-based sciences. This paper thus addresses the following question: why should we take a conclusion deduced from an idealized model to be informative about a target system, given that the proof from which it was derived makes use of falsehoods about the target system?

There are several ways in which scientists and philosophers have tried to lend credence to conclusions derived in this way. Modelers may modify the assumptions in such a way that the dependence of the model-result on the false assumptions is attenuated: a model can be de-idealized (e.g., Peruzzi & Cevolani 2021) or generalized by removing an idealization (Lehtinen 2021, 2022), and one may demonstrate the robustness of the result to false assumptions (e.g., Levins 1966, Weisberg 2006, Kuorikoski et al. 2010). These are all accounts that focus on what modelers themselves can do to cope with falsity. While they are useful in understanding the practices of modeling, these practices share a common epistemic problem: some false assumptions always remain even after de-idealization, generalization, and demonstrations of robustness.

Collin Rice (2019a, b) has recently argued that these strategies for justifying model-based inference (de-idealization, generalization, robustness) are based on assumptions that are not likely to be satisfied by models and their targets. They require that the target be decomposable in such a manner that its relevant features can be isolated from its irrelevant ones. Then, the contributions of the accurate parts of the model must be distinguished from the inaccurate ones, and the successful parts of the model must be mapped to the relevant parts of the target. The inaccurate parts then only concern the irrelevant parts in the real-world system. Rice argues that the decomposition strategy is unlikely to succeed because idealizations are often 'holistic' or 'pervasive' in that they are necessary for the application of useful mathematical techniques. He proposes that models can explain, however, in virtue of the fact that the model and the target may both belong to the same universality class (see also Batterman and Rice 2014, Rice 2020).

Using a formal model from economics, this paper shows that it is possible to usefully decompose a model into its relevant and irrelevant aspects, and demonstrates that a conclusion of interest proven from the model, even though it was proven by the authors using unrealistic assumptions from dif-

ferential calculus, can in fact be deduced solely from the true aspects of the same model, without relying on the idealizations. The method we introduce, which we call 'Soundness Analysis', thus provides a systematic procedure for just the sort of model decomposition that Rice has argued cannot be done: when a Soundness Analysis is successfully performed, it shows that a conclusion of interest can be proven without any false assumptions, even when the relevant causal factors themselves are represented with idealizations, and that the true and false, as well as relevant and irrelevant components in a model, can all be identified.

We are not analyzing what modelers do, whether it is to conduct robustness analyses, de-idealize, generalize, or to establish universality classes. Instead, Soundness Analysis provides a way to show why all these practices succeed, when they do. In this paper, we consider the economic model of Hoff and Stiglitz (2002, 2004), who explain why a certain form of corruption occurred in post-Soviet Russia. The model involves several economic variables which were deemed relevant to the target system and characterizes their relations. However, the model also attributes properties to the target system which it does not have, and the false assumptions play a role in deriving a conclusion that would be difficult to foresee without formal analysis. Namely, that there is a 'multiplier effect', essentially a positive feedback loop, between the relevant economic variables. We demonstrate Soundness Analysis by deriving this conclusion from sound statements about the target system. We do not rely on the proof given by the authors, which uses mathematical techniques from calculus that require idealizations; rather, we provide a logical proof that only relies on the form of the argument to illustrate the structure of the multiplier effect.

The fact that Soundness Analysis provides a strategy for model decomposition raises the issue of how broadly applicable it is. We consider this to be an empirical question: while we believe the Hoff-Stiglitz model is representative of a much broader class of models, and that the possibility of its decomposition lends credence to a wider applicability of Soundness Analysis, we cannot be certain of its usefulness until we attempt similar analyses for other models. In principle, the method can be applied to any setting in which there is a conclusion deduced from a model of a target system, where the model involves assumptions that the modeler knows, *ex ante*, are not true of the target. This is a broad class, which includes both the physical and social sciences, and we take this to be a virtue of Soundness Analysis over other decomposition strategies such as renormalization group methods

(see e.g., Batterman 2002, 2019, Wu 2022), which do not apply to the social sciences. We thus treat the prevalence of decomposability to be subject to empirical investigation via Soundness Analyses: while we agree that not all model results can be derived by such a decomposition, the analysis of this paper shows that some can. Further applications of Soundness Analysis will tell how broadly applicable it is.

The rest of the paper is structured as follows. In Section 2, we discuss how Soundness Analysis relates to the prevailing accounts of modeling epistemology. In Section 3, we introduce Soundness Analysis. In Section 4, we illustrate it by performing a Soundness Analysis on the model of Hoff and Stiglitz. In Section 5, we argue that Soundness Analysis is likely to be more broadly applicable than renormalization group methods.

2 Model Epistemology

This section presents some of the major accounts of model epistemology to clarify the nature of our contribution. It used to be popular to discuss the relationship between a model and the world in terms of similarity instead of truth. For someone like Giere (1988), the reason for proceeding in this manner is precisely the horror of finding falsehoods everywhere in models. Unlike truth, similarity admits of 'degrees and respects', and if one uses this more relaxed notion, there is some hope to find at least some kind of a match between the model and the target. Authors using similarity as the primary relationship between the model and the target then usually proceed by assuming that the inferences from the model to the target are inductive (for an explicit statement, see Sugden 2000, p. 23). This paper seeks to identify the epistemic virtue of theorem-proving in science. Relying on similarity is inadequate for this task because the epistemic virtue of proofs is that they preserve truth values, but proofs do not alter the degree of similarity between a model and its target. From a similarity perspective, an assumption should carry the same epistemic weight as a conclusion that is deduced from a model; thus, similarity or any other induction-based account of model representation cannot explain the epistemic value of deductive reasoning.

Given that Soundness Analysis focuses on extracting a proof that is based solely on true assumptions, it impinges on us to provide an account of the kind of truth we mean. For this purpose, it is helpful to distinguish between idealizations and abstractions. Idealizations are representations that distort

some properties in a target, and abstractions are representations that omit properties that are present in a target (Jones 2005). Soundness Analysis is a method for figuring out whether idealizations can be removed from a model while preserving its conclusion, and it culminates in an abstract representation. What is omitted may be relevant or irrelevant, and when such factors are relevant, the conclusion of the model would be different if the omitted factor were taken into account. Given that Soundness Analysis cannot rule out the possibility of omitting relevant factors, the truths it establishes are always qualified with a ceteris paribus clause. While this may sound like a restrictive presupposition, it is exactly the reverse. It means that Soundness Analysis can be applied even to theorems that are known to be false in the sense that they abstract from causal factors that are known to be relevant, and hence the conclusion of the theorem is at best true in a counterfactual sense. There is, for example, no in-principle obstacle to conducting a Soundness Analysis on the First Theorem of Welfare Economics, or the Hardy-Weinberg Law, even though the conclusions of these theorems are commonly taken to be false about their target systems.

In contrast to our aim of demonstrating soundness, prevailing modeling practices have attempted to mitigate, though not eliminate, the role of false assumptions. Demonstrating the robustness of a model (e.g., Weisberg 2006, Kuorikoski et al. 2010, Lloyd 2015) aims to determine 'whether a result depends on the essentials of the model or the details of the simplifying assumptions' (Levins 1966, p. 20). Weisberg describes the importance of (a true) core structure as follows:

> If a sufficiently heterogeneous set of models for a phenomenon all have [a] common structure, then it is very likely that the real-world phenomenon has a corresponding causal structure. This would allow us to infer that when we observe the robust property in a real system, then it is likely that the core structure is present and that it is giving rise to the property. (Weisberg 2006, p. 739)

The core structure is identified inductively because robustness is a property of a set of models. The basic idea is to tweak the idealizing assumptions of a model and check whether the conclusion of interest is still deductively entailed. From the point of view of the modelers interested in constructing proofs, models consist of various assumptions, $\{A_i\}$, that jointly deductively entail a result R, say $M{=}(A_1A_2A_3A_4A_5){\vdash}\ R$. While the inference is a deductive argument and thereby valid, it is usually not sound because some of the

assumptions are false. Suppose that A_1-A_3 are true but A_4 and $A_5$are false. Robustness may show that the same result can be derived with a model that has replaced a false assumption with another, e.g., $M^r=(A_1A_2A_3A_4A_5')\vdash R$.

De-idealization changes one of the false assumptions into a true one. For example, if A_5^* is true, $M^{r*}=(A_1A_2A_3A_4A_5^*)\vdash R$ provides a de-idealized model compared with M. Finally, one can generalize the result if one can show that $M^g=(A_1A_2A_3A_4)\vdash R$. While each of these model modifications provides some assurance that particular falsehoods are not responsible for the result, they share a common problem, namely, that they don't get rid of all the falsehoods. In our example, A_4 still takes part in deriving the result R. Given that some false assumptions always remain, this kind of predicament creates the need for selecting some particular model elements as 'more important' in generating the conclusion than others, but such assessments are inductive in nature and thus do not justify the soundness of the deduction.

The epistemic importance of robustness has been challenged in several ways. Lisciandra (2017) argued that it is seldom possible to simply replace one auxiliary assumption with another. Instead, what typically happens is that the structure of the model changes, making it more difficult to determine which auxiliaries can be taken to be irrelevant for the result. Odenbaugh and Alexandrova (2011) argued that if it is possible to test every possible auxiliary for robustness, and if it is possible to determine that the true auxiliary is among the ones tested, then robustness does indeed provide epistemic assurance. However, such 'absolute robustness' analysis is seldom possible because it is difficult to define what constitutes the set of all possible auxiliaries.[1] The requirement of being able to determine that the true assumption is among the ones tried reflects a misguided understanding of robustness because the point of testing models for robustness is not to find a true model, but to show the irrelevance of falsities for the robust result. On the other hand, absolute robustness aims at sound arguments just like Soundness Analysis. The difference is that Soundness Analysis does not show the truth of the auxiliaries, but rather the truth of the core structure of the model. Soundness Analysis and most accounts of robustness analysis thus share the aim of establishing the truth of the robust result or the robust theorem by demonstrating the irrelevance of false assumptions[2], but it proceeds by demonstrating the

[1]See, however, Fuller and Schulz (2021) for an example in which this is possible.

[2]The qualification 'most' is needed here because Lehtinen's (2016, 2018) account of indirect confirmation via robustness provides an exception in that it does not necessarily aim to establish the truth of a robust theorem, and its credibility does not hinge on its

soundness of inferences from a core structure rather than by tweaking the false assumptions.

It is thus no accident that models can be robust to tweaking false assumptions: this is a fact that can be explained by the Soundness Analysis because any robustness check which preserves the sound proof structure will preserve the conclusion deduced from it. Once the sound proof structure is extracted, other features of the model can be tweaked arbitrarily without altering the conclusion of interest because they are redundant. Soundness Analysis differs from derivational robustness in requiring that the sound argument structure used to justify a conclusion is, in a precise sense, already entailed by the model. Robustness checks modify the assumptions beyond what the model already entails. Thus, Soundness Analysis can justify a particular proof within the confines of the particular model and its target system, without reference to an ill-defined set of 'independent' and 'sufficiently heterogeneous' models (see also Harris 2021).

3 Soundness Analysis

In this section, we introduce a framework for conceptualizing the relationship between a model and its target system which culminates in a formal definition of Soundness Analysis. To make sense of proof-based reasoning from idealized models, we first review some concepts from mathematical logic, whose subfields of proof theory and model theory are the natural setting for situating our analysis. Since the main ideas can be understood without excessive formalism, we proceed rather informally.

Distinguishing between syntax and semantics will be fundamental in what follows. Consider the formula $\forall x \exists y(x < y)$: if we present this formula as a string of symbols in a formal language, without interpreting the symbols as referring to anything, then we have only specified its syntax. The formula is said to be 'given semantics', when the variables 'x', 'y', and the symbol '$<$' are 'interpreted' as referring to a domain of objects in a 'structure' (or 'model'). For example, we can interpret the variables as referring to the integers, and the symbol '$<$' as a binary order relation, in which case the formula can be interpreted in the structure $(\mathbb{Z},<)$, where $\mathbb{Z}$ is the set of integers, and '$<$' is represented extensionally as a set of ordered pairs, e.g., {(-1,0),(-1,1),(-1,2),...,(0,1),(0,2),(0,3),...} satisfying the axioms of an order

truth.

relation. Note, importantly, that 'interpretation' and 'semantics' are terms of art in this context, and do not fully capture the philosophical notion of semantics as *meaning*. Model-theoretically, the symbol '$<$' is interpreted when its referent in terms of a domain set is specified. Informally, we can say that we 'interpret' this symbol as the 'less than' relation, in which case we would interpret the sentence as *meaning* that 'for all integers x there exists an integer y such that x is less than y'. However, we would then be speaking about interpretation in the 'metalanguage' which we use to speak about the formal language. In practice, it is convenient to allow context to differentiate these senses of 'interpretation', but it is important to keep the distinction in mind.

Given a formal language, which is all finite strings of logical (e.g. $\wedge$, $\vee$, $\neg$) and non-logical (e.g. '$<$') symbols, constructed to meet certain grammatical requirements, the set of non-logical symbols of the language can be organized into a list called a *signature*. A (first order) structure is a tuple $\langle D,\{R_i\}\rangle$, where D is a domain set, and $\{R_i\}$ are the extensions of all the relations from the signature.[3] We will find it convenient to work with 'many sorted' signatures, which have multiple domain 'sorts' (for example, a domain for 'agents' and another domain for 'goods'), and the relevant variables and relations are defined on their respective domains in the usual way.

The distinction between syntax and semantics reflects the division in logic between, respectively, proof theory and model theory. Proof theory studies formal rules of symbol manipulation and inference on the syntax of a language, whereas model theory studies their semantics within structures. The following concepts from logic also respect this distinction: the first represents a syntactic relationship between (sets of) formulas, and the second represents a semantic relationship between a formula and a structure. We say that a sentence $\mathbf{s}$ is 'provable' from another sentence Γ, denoted $\Gamma\vdash\mathbf{s}$, if $\mathbf{s}$ is derivable from Γ following a system of syntactic rules of inference. We say that a sentence $\mathbf{s}$ is 'satisfied by' a model $\mathbf{M}$, denoted $\mathbf{M}\models\mathbf{s}$ if it is true of $\mathbf{M}$. Thus, the sentence $\forall x\exists y(x < y)$ is satisfied by the structure of the integers under the 'less than' order relation, but it is not satisfied by, for example, any structure whose domain is a finite set of numbers, when the relation is interpreted (again, in the metalanguage) as 'less than'.

Model-theoretically, 'truth' is always defined relative to a particular model, via the satisfaction relation. However, since we seek to justify the transfer

[3]Formally, an n-ary relation is a subset of D^n.

of proofs from models to target systems, and since the epistemic value of proofs is their preservation of truth values, we will only consider a sentence to be true if it is satisfied by the target system, treated as a model-theoretic structure.[4] Any model will satisfy some sentences which are also satisfied by the target system (if nothing else, there are exceedingly trivial sentences like 'x=x' or tautologies), and some sentences which are not. For example, the 'common knowledge' assumption in game theoretic models, which can be characterized by an infinite conjunction of knowledge relations ('Alice knows that Bob knows that Alice knows that...'), will not be satisfied by any real-world target system, so cannot be considered true in our framework. The purpose of Soundness Analysis is to derive a conclusion of interest syntactically (here we refer to the provability relation) using only true (in the above sense) premises. The method thus integrates both syntax and semantics: the sentences which serve as true premises are identified semantically (in terms of their satisfaction by both model and target system), but the deductive inference itself is purely syntactic.[5] While there is bound to be contention over which sentences are true of their targets, there are some sentences that are clearly false of their target system, such as the common knowledge sentence in game theory (or any sentence which can only be satisfied by a structure with an infinite domain). Such *ex ante false* sentences cannot feature in any Soundness Analysis. There may be some room for disagreement over what is permitted to be included once these false sentences have been eliminated, but we take this to be a virtue of Soundness Analysis: it forces the theorist to be explicit about what they take to be true about their target.

To help build intuition, conceive of the target system as instantiating a first order structure $\mathbf{W}$ (for 'world') which we do not have full access to. However, we do have *a priori* access to a set of formulas, $\mathbf{T}$, which are true of the target system, that is, $\mathbf{W} \models \mathbf{T}$. A model, $\mathbf{M}$, is a first order structure that has an overlapping vocabulary with the formulas in $\mathbf{T}$. We assume that

[4]Note that our use of model theory does not commit us to a structuralist conception of model representation. We do share with structuralists an emphasis on the relations between objects over the intrinsic natures of the relata; however, structuralists tend to regard the representation relation between model and target as one of morphism, whereas we treat models as representing their targets via sentences they both satisfy. We leave it to future work to explore the nature of this representation relation, and whether and how it relates to structuralism.

[5]For ease of communication, we convey the proof by giving a meta-language interpretation of each inferential step, but since we do not rely on the properties of any domain sets in the proof, it is a syntactic derivation.

a correspondence between the vocabularies in $\mathbf{M}$ and $\mathbf{T}$ can be made in a sufficiently unproblematic way so that it is clear how to translate formulas satisfied by $\mathbf{M}$ into statements about $\mathbf{W}$. For example, an economic model might use 'p' for price, which will have a technical meaning in the context of the model but is naturally associated with prices in the real-world target.

Now consider a formula, $\mathbf{c}$, which is a conclusion mathematically derived from the model, so that $\mathbf{M} \models \mathbf{c}$.[6] The task of Soundness Analysis is to extract from $\mathbf{M}$ a set of 'true premises', $\mathbf{P}$, which syntactically entail $\mathbf{c}$, ($\mathbf{P} \vdash \mathbf{c}$), and are also true, that is, they are elements of $\mathbf{T}$.[7] This brings us to a formal statement of Soundness Analysis:

Definition 1. *A Soundness Analysis of a conclusion formula $\boldsymbol{c}$ satisfied by a model $\boldsymbol{M}$ with respect to a target system $\boldsymbol{W}$ satisfying a set of truths $\boldsymbol{T}$, is a set of 'true premises' $\boldsymbol{P} \subset \boldsymbol{T}$, with $\boldsymbol{c} \notin \boldsymbol{P}$, such that $\boldsymbol{M} \models \boldsymbol{P}$ and $\boldsymbol{P} \vdash \boldsymbol{c}$.*

Treated syntactically, the true premises are uninterpreted formulas, and they may be interpreted in radically different ways in the model and target; however, since the conclusion is derived from these premises syntactically, their interpretation in terms of objects is unnecessary for demonstrating soundness. We take this to be an important finding of Soundness Analysis: although proofs given by scientists are invariably semantic derivations that rely on properties of domains not instantiated in their targets, our analysis shows that it is sometimes possible to extract a sound argument structure that derives the same conclusion while never relying on the intrinsic natures of the objects in the domain. For example, a typical economic model (including the model we consider in the next section) might characterize an agent's 'entrepreneurial ability' by a variable $a \in [0, 1]$. Many of the properties of the real numbers are not satisfied by any actual economic system, but there is indeed some (perhaps partial) order relation on entrepreneurial abilities, just as there is an order relation on the real numbers between zero and one. Although the semantics of the economic model will characterize the (binary) order relation as a subset of $[0, 1]^2$ satisfying the properties of a linear order, and the semantics of the target system will characterize the binary relation

[6] Since $\mathbf{c}$ is provable from formulas satisfied by $\mathbf{M}$, it too is satisfied by $\mathbf{M}$. We are here taking for granted Hilbert's Thesis that every mathematical proof can be stated as a formal, logical proof. This is basically to assume that mathematical proofs preserve truth values (See Boolos et al. 2007, p. 185).

[7] Note that the formulas in $\mathbf{T}$ must be consistent since they are satisfied by $\mathbf{W}$, but they need not be deductively closed.

over a set of people which are at best partially ordered according to their entrepreneurial abilities, these differences in semantics are irrelevant from the perspective of the syntactic derivation. As long as the derivation only relies on true premises, which are in this example the syntactic properties of the order relation, theorems from the model can be transferred to the target as sound arguments for the conclusion of interest.[8] The fact that differences in formal semantics are irrelevant for transferring proofs from a model to its target should alleviate what Odenbaugh (2021) calls 'Hughes' worry' about the inability of mathematical objects to have physical properties. Once we distinguish between syntax and semantics, we can see that physical and mathematical systems can share syntactic properties which have nontrivial deductive implications about their shared behavior.

Thus, given a model and a conclusion derived from it, Soundness Analysis proceeds in two steps. First, one must extract premises, $\mathbf{P}$, satisfied by the model which are true of (satisfied by) the target system. Then, one must show that these formulas syntactically entail the conclusion. In practice, finding $\mathbf{P}$ involves abstracting the original model descriptions in such a way that only truths remain. It is helpful to organize the syntactic vocabulary of the Soundness Analysis into a many-sorted signature because then Soundness Analysis can be interpreted as characterizing the names of objects and relations present in both the model and target system, as well as the syntactic proof structure which entails the conclusion of interest. Thus, Soundness Analysis effectively takes a conclusion from a proof and asks: what are the names of the objects and relations in the model which are present in the target, and what can be derived from their syntactic properties alone? Then, instead of taking a model literally as explaining a conclusion using falsehoods, a model can be 'reconstructed' in terms of a sound deductive structure thatFentails the conclusion of interest.[9]

We demonstrate Soundness Analysis with an applied game theoretic model from political economy, but the method can in principle be attempted for any explanatory model which derives conclusions about a target system deductively. The selected model is representative in the sense that its theorems

[8]Along the way, it is convenient to interpret the syntactic derivation in the metalanguage, for example by referring to the order relation as a 'greater-than' relation, but this is done for ease of exposition rather than out of necessity.

[9]Note that this will not always be possible. One might not be able to find an appropriate set of true premises which entail a conclusion of interest. It takes methodological work to demonstrate this via Soundness Analysis.

purport to offer insights about a real-world phenomenon, even though the model is highly idealized and the proofs make use of these idealizations.

4 Application

In this section, we illustrate Soundness Analysis on a theorem from the economic model of Karla Hoff and Joseph Stiglitz (2002; 2004). Their model purports to shed light on the relationship between private property rights and respect for law in Russia in the immediate aftermath of the fall of the Soviet Union. In that transitionary rebuilding period, a prominent view held that to encourage respect for the rule of law, transferring ownership from the state to private citizens would create demand for private property rights and rule of law procedures that respect these rights. The intuition was that establishing private property rights would enhance respect for the rule of law because those with private property would have a greater influence on the government's procedures than those without it, and the former would have an incentive to urge rule of law practices upon the government.

Hoff and Stiglitz' model explains why, despite the plausibility of the above story, privatization did not lead to substantial respect for the rule of law. Their insight was that increasing private property rights in a state of low respect for the rule of law can increase the incentives of business owners to 'steal' from their businesses rather than invest in them. We begin by presenting their model and then perform a Soundness Analysis on their main 'surprising' result.

There is a continuum of agents that are defined by their 'asset-stripping ability', $\theta \epsilon [0,1]$, so that those with higher values of θ have the higher asset-stripping ability. Each agent owns a firm. There are two time periods. In the first period, agents decide whether to invest in their firm (build value) or strip assets from it. Initially, there is no rule of law, and agents who build value demand rule of law reforms whereas agents who strip assets 'vote' for the status quo. The status quo block is fraction x, and the constituency for the rule of law reform is a fraction 1-x. An agent of type θ who chooses to strip assets in period 1 from their firm receives a payoff of $s(\theta)=f+\theta$, where f is the amount the firm produces in period 1. Agents who invest in period 1 obtain period 2 value $V_j=f+g-I_j$, where g is the growth of the asset, and I_j is the cost of investment, which depends on the status of the rule of law: j takes either the value 'law', L, or 'no law', N. It is assumed that the cost

of investment under 'no law' is greater than the cost of investment under 'law': $I_N > I_L$. It is also assumed that more than half the agents are better off building value under the rule of law than stripping assets: g-$I_L > \theta_{median}$, the intuition being that a majority of agents would build value if the rule of law were a certainty.

The probability of establishing the rule of law, π, is a differentiable function of the constituency: $\pi = \pi(\mathrm{x})$ (recall x is the constituency for 'no law'), and since the probability of establishing the rule of law is decreasing in the constituency for 'no law', this function has a negative first derivative. Given a constituency for status quo x, the 'threshold' type, θ^*, above which agents vote for the status quo (and below which they demand the rule of law) is given by the 'switch line':

$$\theta^*(\pi(x)) = g - \pi(x)I_L - [1 - \pi(x)]I_N \tag{1}$$

A cumulative distribution function $\mathrm{H}(\theta^*(\pi(\mathrm{x})))$ characterizes the fraction of agents who support the rule of law (therefore 1-$\mathrm{H}(\theta^*(\pi(\mathrm{x})))$ support the status quo). Given this setting, an equilibrium is defined in terms of a constituency for the status quo:

Definition 2. *An equilibrium is a constituency x such that x=1-H($\theta^*(\pi(x))$).*

Hoff and Stiglitz derive two propositions from this setup. The first establishes conditions for the existence of multiple equilibria; in particular, the possibility of 'rule of law' equilibria as well as 'status quo' equilibria. They interpret this conclusion as follows:

> The model sheds light on the debate about rapid privatization. ...Those who have an advantage in asset-stripping relative to wealth creation may also have an advantage in converting corporate and social assets to private use, and accordingly will not support the rule of law even when they themselves have assets to protect. As the oligarch Boris Berezovsky might have said, Why create when you can steal? Our analysis suggests that...building value may be rational and stripping assets may be rational - but that unfortunately there can exist an equilibrium in which the latter prediction on this issue is the one that seems to have emerged.
>
> (Hoff and Stiglitz 2002, p. 14)

We have here a clear case of a model with falsehoods being used to deduce implications for a real-world phenomenon. Despite this, the intuition is entirely familiar: the situation is a coordination problem, and they tend to have both 'cooperative' as well as 'non-cooperative' equilibria. Although the assumptions of the model clearly involve falsehoods, the claim that rule of law reforms presented a coordination problem is plausible.

Perhaps the authors could have done without the model altogether if this had been their only conclusion; however, they derive a second proposition from their setup which is far from obvious, and seems to only be accessible from a mathematical characterization of the target system:

> Proposition 2: There is a social multiplier. Parameter changes have both a direct effect on the demand for the rule of law...and an indirect effect of the same sign. (Hoff and Stiglitz 2002, p. 15)

In our Soundness Analysis, we seek to derive this proposition from true premises; that is, formulas satisfied by the model which are also true about the target system. The possibility of doing so gives epistemic license for taking Proposition 2 to be an insight into the behavior of the target system, even though it is derived using falsehoods in the model.

The first step in characterizing the sound argument structure is determining which objects play an essential role in deriving the conclusion. For example, in the model of Hoff and Stiglitz, individual agents 'vote' for rule of law reforms, but the votes of individual agents do not play a role in the conclusion, only the sizes of constituencies do. Therefore, Soundness Analysis includes a domain 'sort'[10] of constituency sizes, making no reference to individual agents. Soundness Analysis, framed in a many-sorted logic, involves a statement of these 'essential' domains, as well as relation symbols[11] and the syntactic properties they satisfy in both the model and target. Consider now the following sorts with their meta-language interpretations:

1. X is to be interpreted as a domain of constituency sizes, with variables 'x'.

[10] A sort does not specify the objects of the domain, so it is a syntactic characterization of the domain.

[11] For convenience, we will use functions, but this is without loss of generality as every n-ary function to a given codomain can be represented as an n+1-ary relation on these domains.

2. P is to be interpreted as a domain of probabilities (of establishing the rule of law), with variables 'p'.

3. S is to be interpreted as a domain of asset-stripping abilities, with variables 's'.

These sorts are, from a syntactic perspective, uninterpreted (the objects of their domains are left unspecified), but here they are named suggestively based on the objects that they will be interpreted to denote (in the meta-language). Recall why this is important: we seek to characterize true premises for a sound argument, and this requires us to characterize only the syntactic properties which are satisfied by both model and target.

It is not necessary to place much syntactic structure on these domains. For example, it will not even be required that the domain of probabilities satisfies the Kolmogorov Axioms. We will only require that all the domains are totally ordered, so that, for example, the extension of the probability sort in the target system could be as simple as P=(low, medium, high). The following total orders are binary relation symbols on their respective sorts:

1. $<^X$ is a total order on the domain of constituency sizes.

2. $<^P$ is a total order on the domain of probabilities.

3. $<^S$ is a total order on the domain of asset-stripping abilities.

In addition to these binary relation symbols, we introduce three function symbols (the function symbols are interpreted as functions in the meta-language).

1. π_L:X$\rightarrow$P outputs, for a given constituency size x, the probability of establishing the rule of law. This function is assumed to be decreasing, ie. $x^*<^X x \rightarrow \pi_L(x)<^P \pi_L(x^*)$. 'The larger the constituency who favor the status quo, the lower the probability of establishing rule of law.'

2. Θ:P$\rightarrow$S outputs, for a given probability value, an asset-stripping ability. This function is assumed to be increasing, ie. $p^*<^P p \rightarrow \Theta(p^*)<^S \Theta(p)$. 'The greater the probability of establishing rule of law, the greater the asset-stripping ability of the supporters of the status quo.'

3. F:S$\rightarrow$X outputs, for a given asset-stripping ability, a constituency for status quo, x. This function is assumed to be decreasing, ie. $s^*<^S s \rightarrow F(s)<^X F(s^*)$. 'The greater the asset-stripping ability of the supporters of the status quo, the lower the constituency for the status quo.'

The many-sorted signature can be summarized by the tuple $\langle$(X, P, S),($<^X$,$<^P$,$<^S$), (π_L,Θ, F)$\rangle$.

These formulas are all the true premises we will need. They express relationships between economic variables in the target system which conform to our *a priori* understanding of the target system. First, it stands to reason that increasing the constituency for a political outcome increases the probability of its occurrence. Second, if voters refuse rule-of-law reforms based on their asset-stripping abilities, then when there is a high probability of rule of law reforms, only those with high asset-stripping abilities will refuse them. Finally, if the supporters of the status quo are only those with high asset stripping abilities, then they will comprise a small constituency.

Having extracted these true premises, Soundness Analysis requires two demonstrations: first, the true premises must be satisfied by the model, and second, the true premises must syntactically entail the conclusion. The first requirement is straightforward:

1. Sorts X, P, and S are all modeled by Hoff and Stiglitz as [0,1].

2. The binary relations which totally order X, P, and S are the usual order on real numbers.

3. The π_L function of the Soundness Analysis corresponds to π:[0,1]$\rightarrow$[0,1] in the original model. Recall that this function is decreasing in the model.

4. The Θ function of the Soundness Analysis corresponds to the switch line θ^*:[0,1]$\rightarrow$[0,1] in the original model. Note that since $I_N > I_L$ this function is increasing in the probability of establishing law.

5. The F function corresponds to 1-H($\theta^*(\pi(\text{x}))$) in the original model. Note that this function is decreasing in the cutoff θ^*.

For the second requirement, consider the effect of an exogenous increase in the probability of establishing law at a given constituency x, i.e., a shift from π_L(x)=p to π_L(x)=p*, where p$<^P$p*. Then, since Θ is increasing, we have Θ(p)$<^S\Theta$(p*), and since F is decreasing, we have F(Θ(p))$>^X$F(Θ(p*)), which for convenience we can write as x$>^X$x*. This is the first-order 'direct' effect on demand: an exogenous increase in the probability of establishing law leads to a decrease in the constituency for the status quo.

However, given this first-order decrease in demand for the status quo, there is a second-order 'indirect' effect of the same sign because, at the new probability of the rule of law, $\pi_L(\text{x})=\text{p}^*$, the decrease from x to x* leads (since π_L is decreasing in x) to $\pi_L(\text{x})<^P\pi_L(\text{x}^*)$. Since we had $\pi_L(\text{x})=\text{p}^*$ it follows that $\text{p}^*<^P\pi_L(x^*)$. Thus, the probability of establishing law is further increased (by the first-order decrease in demand for the status quo), which has a second-order effect of the same sign on demand for the status quo (it is further decreased), and the argument above repeats, producing the desired multiplier effect.[12]

Hoff and Stiglitz arrive at this multiplier effect by relying on the specific features of their model - their assumption that the relevant functions are differentiable enables them to use tools from calculus such as the implicit function theorem to derive the conclusion. However, as we have shown, the effect follows from true premises which do not refer to the falsehoods of their model. Furthermore, any structure satisfying these premises, as a matter of logic, will have a multiplier effect; the model of Hoff and Stiglitz is just one member of this class of models. We take this abstractness to be one of the benefits of a syntactic characterization of deductive reasoning. Radically different systems may have the same syntactic structure entailing a multiplier effect just as a variety of unrelated physical systems can be described in terms of the mathematics of harmonic oscillators. It is only by interpreting the syntax in a meta-language that we think of them as referring to particular real-world entities.

The substantive question, then, is whether the target system is appropriately conceived of as a member of this model class; that is, are the extracted premises indeed true? For example, in the immediate aftermath of the fall of the Soviet Union, was there really a relationship between asset-stripping ability and the probability of establishing the rule of law obeying the properties of the Θ function symbol?[13] Was it really the case that the probability of establishing the rule of law was decreasing in the constituency for the status quo? The logic of the Hoff and Stiglitz model assumes the existence of these relationships, and they present some evidence in their support. By performing a Soundness Analysis, the question of how a model involving falsehoods can represent a real-world target system has been transformed

[12]A similar argument can be made to deduce a multiplier effect from a shift in the asset-stripping ability cutoff.

[13]Note that we refer to this as a function symbol rather than a function because we have only given it a syntactic characterization.

into one of whether realistic relationships between economic variables were indeed present and relevant in the historical target system. We can now see that the falsehoods of their model (for example, the continuum of agents, each owning a firm, satisfying game-theoretic rationality assumptions, and so on) play no role in the model entailing a multiplier effect. Whether the assumptions included in the Soundness Analysis are indeed true is an empirical question. Given the evidence for these truths, however, we are licensed in believing (as ceteris paribus claims) conclusions they deductively entail via the argument structure of the Soundness Analysis.

5 Decomposability and Explanation

Soundness Analysis and university classes both involve systematic methods for decomposing a model and its target into relevant and irrelevant components, but their consequences for the debate on the role of veridicality in explanation (e.g., Pincock 2023) are different. In this section, we explore Rice's non-decomposability claim and universality classes in some more detail, in light of the possibility of Soundness Analysis.

Minimal models, according to Batterman and Rice (2014), are "thoroughgoing caricatures of real systems'' whose explanatory power does not depend on their "representational accuracy'' (p. 350) that are "used to explain patterns of macroscopic behavior across systems that are heterogeneous at small scales" (p. 349). Batterman and Rice argue against "shared features" accounts, which posit that properties shared by a model and its target explain the target's behavior. They argue instead that models explain in virtue of the fact that the model and the target may both belong to the same universality class. Then, given that belonging to the same universality class is sufficient for explaining why a model and target exhibit the same macro behaviors, all the falsities in the model concerning further details become irrelevant.

Rice's argument for non-decomposability goes as follows (Rice 2019a, 2020, 2021). Models use idealizations to be able to take advantage of various mathematical techniques. The standard shared-features account claims that one can show that the idealizations are irrelevant. But this account faces the "serious problem" that idealizations often make positive epistemic contributions by directly distorting relevant causes of the target phenomenon (2021, p. 38, see also Grï¿œne-Yanoff 2009). Given that the relevant causes are represented with idealizations, it is impossible to distinguish between

the irrelevant and relevant falsities, and thus impossible to show that the idealizations only concern irrelevant features.

Rice seems to think that since the relevant factors are represented with idealizations, this means that one cannot separate the idealizations from the true representations. Such a view overlooks the fact that mathematical functional forms typically entail assumptions at several levels of abstraction simultaneously, such that the specific ones entail the abstract ones, but not vice versa (Rol 2008, Strevens 2008, Lehtinen 2022, see also Saatsi 2013). A successful Soundness Analysis abstracts from a given model in the sense that it strips away various assumptions that concern specificities that are ultimately not necessary for the result.

A proposition, p*, is more abstract than a proposition p, if p* describes the same target as p with less content in the sense that it is less detailed (Levy 2018). For example, the proposition that 'demand is inversely proportional to price' is more abstract than the proposition that 'demand is inversely proportional to price via the equation D=1/p'. Although models invariably characterize propositions that are false of their targets, our analysis shows that it is also possible to extract propositions such as these which are satisfied by the model and its target. Thus, Soundness Analysis is squarely a shared-features account.

The explanandum phenomena for universality classes are not particular systems but rather classes of them, in that they account for a large number of particular systems (McKenna 2021). Soundness Analysis is framed in terms of a particular model and its target, but its method can also explain features of large numbers of particular systems. Recall that Soundness Analysis is always relative to a given model conclusion, which corresponds to the explanandum phenomenon. The truths extracted from the model that deductively entail the conclusion of interest will apply to a class of target systems. For example, the structure of the multiplier effect, which we characterized syntactically in the previous section, will be satisfied by systems radically different from post-Soviet Russia in their details, but which all belong to the same model class. A model class of a set of sentences is the set of models which satisfy those sentences. Thus, the sound 'proof structure' of a Soundness Analysis explains not only a particular target system's phenomenon but simultaneously characterizes the structure of a class of systems that all exhibit that same phenomenon.

Soundness Analysis shows that idealized mathematical representations are useful to the extent that they characterize true propositions about their

targets, and models provide sound arguments for their conclusions to the extent that their conclusions can be derived from these truths, independently of the idealizations used to characterize them. Thus, when Soundness Analysis succeeds, it shows that the model result is decomposable: the result is explained purely by the true components of the model. Unlike the eliminative procedure for finding causal difference-makers in Strevens (2008), the method of extracting shared syntactic properties between a model and its target justifies deductive inferences from idealized models.

Insofar as one is concerned with establishing that the result of interest depends on true assumptions, the original scientific model must be partially true (Yablo 2014, Levy 2015, Pincock 2023) with respect to the truths that Soundness Analysis extracts. Thus, if the extracted assumptions are true, and if they are sufficient for entailing the result of interest, Soundness Analysis can demonstrate this. However, the method itself does not require the truth of the extracted properties. Furthermore, if the original model result is true only ceteris paribus, so is the abstracted Soundness Analysis result. This is why, when we talk about truths, they are always to be interpreted in this ceteris paribus sense.

The possibility of Soundness Analysis shows that pace Rice, representing relevant factors with idealizations does not in itself preclude a shared features account of explanation. The tenability of shared features accounts depends on the degree to which Soundness Analysis can illuminate a variety of scientific models. Rice's non-decomposability claim is thus ultimately an empirical one, and it is to be evaluated by looking at various scientific cases. We now argue that, if one takes all scientific models as potentially relevant test cases, the idea that minimal models explain by appealing to universality classes is likely to be much less broadly applicable than Soundness Analysis.

There is a strict and broad interpretation of what it means to belong to the same universality class. In the strict interpretation, systems that 'flow toward the same fixed point' are said to be in the same universality class. The systems must be characterized by the same critical exponent, in other words, a common parameter describes several microphysically different systems in some circumstances. The broad interpretation posits that systems belong to the same universality class if they display similar behaviors despite differences in their physical features (e.g., Kadanoff 2013, Rice 2020). To put it differently, systems within the same universality class display the same behaviors in multiply realized ways (see Batterman 2000). Rice's (2019a, 2020, 2021) non-decomposability argument is intimately coupled with the

idea that one can use universality classes in a broad sense to explain when decomposability fails.

Just about all results in economics and the social sciences concern multiply realizable properties with different underlying causal mechanisms. For example, the structure of the multiplier effect in the Hoff-Stiglitz model arises from constituencies for rule-of-law reforms, but multiplier effects also occur when central banks stimulate markets, via entirely different causal mechanisms. Presumably, given the multiple realizability of the multiplier effect, this would be a prime case study for applying universality classes. However, a universality class in the strict sense cannot explain the multiple realizability of the multiplier effect. Showing that a system belongs to a universality class in the strict sense requires renormalization group methods which have limited applicability outside of physics (see Rice 2021, 2022 and the references therein for cases that he has discussed). We are aware of only one effort that applies such methods in economics: the econophysics of finance. But as Jhun et al. (2018) argue, the relevant macro behaviors ('critical' market crashes) do not form a universality class in the strict sense of renormalization group physics.

Consider now whether universality classes in the weak sense can explain the multiplier effect. Rice (2021, Ch. 6) says that universality simply means the stability of certain patterns or behaviors across systems that are heterogeneous in their features, and he also calls this the 'fact of universality'. The fact of universality is thus just another way of saying that something is multiply realizable. We agree with Rice (2021) that belonging to a universality class may explain even when one cannot provide a complete explanation of universality itself, and even when one does not know exactly which systems belong to a universality class. However, merely pointing to the fact that there is a multiplier effect does not explain why there are such effects in these contexts, nor why they can be multiply realizable: description is not explanation. To explain the multiple realizability as well as the common patterns themselves, one has to be able to explain why the differences in the systems are irrelevant.

How, then is the irrelevance of the details demonstrated in the case of broad universality? To show that a system belongs to a universality class in the broad sense, Batterman and Rice (2014) argued that "delimiting the universality class" for Fisher's 1:1 sex ratio result was based on showing that the result is robust. But then, delimiting the broad universality class was achieved by a method, robustness, that is based on decomposition. From this

perspective, it should not be surprising that Soundness Analysis may delimit a universality class in the broad sense: demonstrating that the multiplier effect can be derived with Soundness Analysis in the contexts of rule-of-law reforms with premises that are likely to be satisfied by a variety of economic systems, shows that the multiplier effect is multiply realizable, and thus that they presumably belong to the same universality class. But then, clearly, the fact of universality does not explain in these cases; they are explained by showing that there is an underlying abstract structure entailing their occurrence.[14]

Multiple realizability is thus ubiquitous in the special sciences, but the methods of explaining it require decomposability. Thus, if Rice's claim about pervasive non-decomposability is correct, in sciences that cannot apply renormalization group methods, it becomes a mystery how multiply realized properties and behaviors can be explained in the first place. In contrast, Soundness Analysis provides an explanation of multiple realizability in terms of model classes. Although a multiplier effect may arise in a variety of target systems, each with its own causal difference-makers, these systems share a precisely definable logical structure: they belong to the class of models which satisfy the syntactic formulas entailing the effect. Soundness Analysis thus accounts for multiple realizability with a shared-features account of explanation, and due to its applicability in the special sciences, it has a broader scope than universality class methods.

6 Conclusions

Soundness Analysis is a novel method for extracting from a model a sound deductive structure that entails a conclusion of interest. Its methodological importance lies in the fact that it provides a systematic method of demonstrating that a result only depends on true assumptions. When Soundness Analysis succeeds, it does so because the model result only depends on true assumptions that concern abstract properties rather than the idealizations concerning the details. We demonstrated the method for a particular case study (the model of Hoff and Stiglitz), but the method can be attempted for any deduced consequence of an idealized model. On the other hand, it is

[14]Several authors have already argued that using the strict universality classes reduces to a common features account (Lange 2015, Reutlinger 2017, Povich 2018).

obvious that many results will not withstand a Soundness Analysis simply because they actually depend on false assumptions.

If a model is successfully generalized or de-idealized, or if it is shown to be robust with respect to an unrealistic assumption, Soundness Analysis can explain why these practices may succeed. It does so by showing that a result is a syntactic consequence of a deeper, shared structure in a set of models: all members of the relevant model class will entail the same conclusion. Soundness Analysis thus presents a method for justifying the scientific practice of giving epistemic credence to conclusions deduced from idealized models. That being said, from the perspective of Soundness Analysis, belief in a conclusion derived from an idealized model is grounded in the belief that a Soundness Analysis could be performed on that conclusion.

The true premises extracted by a Soundness Analysis define a model class that satisfies those premises. The syntactic proof of a conclusion of interest from these premises explains how the result occurs in multiply realized, yet 'structurally similar' ways. Furthermore, Soundness Analysis is a tool for quasi-empirically evaluating Rice's non-decomposability claim: whether the various results and theorems that modelers have found illuminating can be justified by a Soundness Analysis is a matter for future applications of the method. Finally, the example of a successful Soundness Analysis on the Hoff and Stiglitz model establishes that, despite the widespread use of idealizations in scientific practice, the idea of a 'true core structure' of a model is not an impossible dream, but rather something that can be demonstrated. This means that Rice's non-decomposability claim is empirical rather than conceptual: there are no in-principle obstacles to decomposing a model result, even when the relevant factors are expressed with idealizations in the original model.

7 References

1. Batterman, Robert (2000): "Multiple realizability and universality", British Journal for the Philosophy of Science 51(1): 115-145.

2. — (2002): The devil in the details: asymptotic reasoning in explanation, reduction, and emergence. Oxford: Oxford University Press.

3. — (2019): "Universality and RG explanations", Perspectives on Science 27(1): 26-47.

4. Batterman, Robert and Collin C. Rice (2014): "Minimal Model Explanations", Philosophy of Science 81(3): 349-376.

5. Boolos, George. Burgess, John. Jeffrey, Richard (2007): Computability and logic (5th edition). Cambridge University Press.

6. Cartwright, Nancy (2007): "The vanity of rigour in economics: theoretical models and Galilean experiments", in Hunting Causes, Cambridge University Press, pp. 217-235.

7. Franklin, Alexander (2018): "On the renormalization group explanation of universality", Philosophy of Science 85(2): 225-248.

8. Frigg, Roman (2022): Models and Theories. Cheltenham:Routledge.

9. Fuller, Gareth, and Armin Schulz (2021): "Idealizations and Partitions: A Defense of Robustness Analysis", European Journal for Philosophy of Science 11(4).

10. Giere, Ronald (1988): Explaining science: a cognitive approach. Chicago: University of Chicago Press.

11. Grï¿œne-Yanoff, Till (2009): "Learning from minimal economic models", Erkenntnis 70(1): 81-99.

12. Harris, Margherita (2021): "The epistemic value of independent lies: false analogies and equivocations", Synthese.

13. Hoff, Karla. Stiglitz, Joseph (2002). After the Big Bang? Obstacles to the Emergence of the Rule of Law in Post-Communist Societies. NBER WP 9282.

14. — (2004). After the Big Bang? Obstacles to the Emergence of the Rule of Law in Post-Communist Societies. American Economic Review 94: 753-763.

15. Jhun, Jennifer, Patricia Palacios, and James Owen Weatherall (2018): "Market crashes as critical phenomena? Explanation, idealization, and universality in econophysics", Synthese 195(10): 4477-4505.

16. Jones, Martin (2005): "Idealization and abstraction: A framework", Poznan Studies in the Philosophy of the Sciences and the Humanities 86: 173-218.

17. Kadanoff, Leo (2013): "Theories of Matter: Infinities and Renormalization", in The Oxford Handbook of Philosophy of Physics: Oxford University Press.

18. Knuuttila, Tarja (2009): "Isolating representations vs. credible fictions? Economic modelling in theory and practice", Erkenntnis 70(1): 59-80.

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

20. Lange, Marc (2015): "On "Minimal Model Explanations": A Reply to Batterman and Rice", Philosophy of Science 82(2): 292-305.

21. Lehtinen, Aki (2016): "Allocating confirmation with derivational robustness", Philosophical Studies 173: 2487-2509.

22. — (2018): "Derivational Robustness and Indirect Confirmation", Erkenntnis 83(3): 539-576.

23. — (2021): "The epistemic benefits of generalisation in modelling I: Systems and applicability", Synthese 199(3-4): 10343-10370.

24. — (2022): "The epistemic benefits of generalisation in modelling II: Expressive power and abstraction", Synthese.

25. Levins, Richard (1966): "The strategy of model building in population biology", American Scientist 54(4): 421-431.

26. Levy, Arnon (2015): "Modeling without models", Philosophical Studies 172(3): 781-798.

27. — (2018): Idealization and abstraction: refining the distinction. Synthese:1-18.

28. Lisciandra, Chiara (2017): "Robustness analysis and tractability in modeling", European Journal for Philosophy of Science 7: 79-95.

29. Lloyd, Elisabeth A. (2015). "Model robustness as a confirmatory virtue: The case of climate science". Studies in History and Philosophy of Science Part A. 49:58-68.

30. McKenna, Travis (2021): "Lange on Minimal Model Explanations: A Defense of Batterman and Rice", Philosophy of Science 88(4): 731-741.

31. Odenbaugh, J. 2011. True lies: Realism, robustness, and models. Philosophy of Science 78 (5):1177-1188.

32. Odenbaugh, Jay (2021): Models, models models: a deflationary view. Synthese. 198 (4).

33. Peruzzi, Edoardo, and Cevolani, Gustavo. (2021). Defending Deidealization in Economic Modeling: A Case Study. Philosophy of the Social Sciences.52 (1-2):25-52.

34. Pincock, Christopher (2023). A defense of truth as a necessary condition on scientific explanation. Erkenntnis 88:621-640.

35. Povich, Mark (2018): "Minimal Models and the Generalized Ontic Conception of Scientific Explanation", The British Journal for the Philosophy of Science 69(1): 117-137.

36. Reutlinger, Alexander (2017): Do Renormalization Group Explanations Conform to the Commonality Strategy? Journal for General Philosophy of Science 48 (1):143-150.

37. Rice, Collin (2019a): "Models Don't Decompose That Way: A Holistic View of Idealized Models", The British journal for the philosophy of science 70(1): 179-208.

38. — (2019b): "Universality and the Problem of Inconsistent Models", in Understanding Perspectivism: Scientific Challenges and Methodological Prospects, edited by Michela Massimi and C. D. McCoy. London: Routledge,

39. — (2020): "Universality and Modeling Limiting Behaviors", Philosophy of Science 87(5): 829-840.

40. — (2021): Leveraging distortions: Explanation, idealization, and universality in science. Cambridge, Massachusetts: The MIT Press.

41. — (2022): Modeling multiscale patterns: active matter, minimal models, and explanatory autonomy. Synthese 200 (6):432.

42. Rol, M. 2008. Idealization, abstraction, and the policy relevance of economic theories. Journal of Economic Methodology 15 (1):69-97.

43. Saatsi, Juha 2013. Idealized models as inferentially veridical representations: A conceptual framework. In Models, simulations, and representations: Routledge, 252-267.

44. Strevens, Michael (2008): Depth: An account of scientific explanation. Cambridge, Mass. ; London: Harvard University Press.

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

46. Weisberg, Michael. (2006): Robustness analysis. Philosophy of Science, 73, 730-742.

47. Wu, Jingyi (2021): "Explaining universality: infinite limit systems in the renormalization group method", Synthese 199(5): 14897-14930.

48. Yablo, Stephen (2014): Aboutness. Princeton, Oxford: Princeton University Press.