Philosophical Studies
https://doi.org/10.1007/s11098-025-02384-y

Stabilising understanding

Roman Frigg[1] · **James Nguyen**[2]

Accepted: 12 July 2025

Abstract
We present an account of how idealised models provide scientific understanding that is based on the notion of stability: a model provides understanding of a target behaviour when both the model and the target's perfect model are in a class of models over which that behaviour is stable. The class is characterised in terms of what we call the model's noetic core, which contains the features that are indispensable to both the model's and the target's behaviour. The account is factivist because it insists that models must get those aspects of the target that it aims to understand right, but it disagrees with extant factivist accounts about how models achieve this.

Keywords Understanding · Explanation · Scientific models · Stability · Idealisation · Robustness · Renormalization · Universality · Stability

1 Introduction

Successful science doesn't just produce representations of the world, those representations are the means through which we *understand* it. Many fields achieve this goal through the construction of models. This raises the question: how do models provide understanding of their target systems?

This has sparked a lively debate about the nature of scientific understanding.[1] A central concern is that most, if not all, models involve idealisations: falsehoods, at least when taken literally. A ski slope is not a frictionless plane; a crystal doesn't

[1] The contributions to Lawler, Khalifa and Shech (2023) document both the state of play in the debate and the trajectory that it followed in the past.

✉ Roman Frigg
r.p.frigg@lse.ac.uk

James Nguyen
james.nguyen@philosophy.su.se

[1] Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, London, UK

[2] Department of Philosophy, Stockholm University, Stockholm, Sweden

Published online: 09 August 2025

consist of an infinite number of molecules; and economic agents aren't perfectly rational. And yet models that represent them as such seem to provide understanding. But how can such distortions do this? Views on how to answer this question are divided. *Factivists* insist that understanding is factive and embark on the project of separating models' veridical and non-veridical parts, insisting that only the former contribute to understanding. *Non-factivists* demur and insist that radical departures from the truth don't impede understanding and should be embraced rather than excised: models can provide understanding irrespective of being literally false.[2]

We take issues with both sides, but for different reasons. We argue that non-factivism faces a dilemma: either it is too permissive to account for how models provide understanding, or, if amended to avoid that charge, it collapses into factivism. We argue for this with a thought experiment inspired by the history of physics which shows that unless a model gets those aspects of the target that we aim to understand right, it fails to provide understanding. This places us in the factivist camp. However, we object to existing articulation of factivism, particularly with respect to how to accommodate the idealised aspects of scientific models within a factivist framework. As we will see, the most promising attempts typically deal with idealisation via a comparison between an idealised model and some more veridical model, which requires that the latter, if not accessible in practice, is at least the sort of thing to which an idealised model can be compared to in principle.

Our strategy for handling idealisation eschews such comparisons.[3] The account we propose is based on the notion of stability. *In nuce*, a model provides understanding of some target behaviour if both the model and the target's "perfect model" are in a class of models over which that behaviour is stable. The class is characterised in terms of what we call the model's "noetic core", which contains the features that are indispensable to both the model's and the target's behaviour. The understanding resulting from the model is then characterised both in terms of the features that are common to models in that class and the features that vary across that class, which correspond to features that don't make a difference to that behaviour.

As we will see, this way of characterising how idealised models provide understanding complements existing discussions of "universality" explanations and the understanding they are associated with (Batterman & Rice, 2014; Rice, 2021). Our discussion refines the structure of how this understanding emerges, and clarifies how the "accurate" and "inaccurate" aspects of those models contribute to that understanding without abandoning accurate representation as a condition on understanding. In doing so, our approach is guided by the idea that the way(s) in which a model provides understanding should reflect the methods and techniques used by scientists when reasoning with their models. As such, the resulting account has the virtue that it unifies a number of styles of scientific reasoning that are widely used across

[2] This divide is also described as one between "veristists" and "non-veritists".

[3] It's also worth noting that our focus here is primarily on the factivist vs. non-factivist issue as regards idealisation. We remain noncommittal regarding other dividing lines in the landscape of understanding, for example the precise role played by "intelligibility" (de Regt 2017) or "grasp" (Strevens 2024). We touch on this below in Sect. 6.

different scientific disciplines, but are the subject matter of separate and disjointed discussions. We bring them together under one umbrella and explain how they contribute to understanding in the sense we propose. In addition to universality, the styles of reasoning we have in mind are robustness analysis (Weisberg, 2006) and the principle of stability (Fletcher, 2020).

Throughout the paper, we employ the following terminology.[4] The *object of understanding* is the thing, typically a part or aspect of the natural or social world, that stands in need of being understood. The *basis of understanding* (alternatively: *source* or *vehicle*) is the thing that provides the understanding. Something that provides understanding is said to be *noetic*. These notions correspond to *explanandum*, *explanans*, and *explanatory* in the context of discussions of scientific explanation.

The ideal gas has become the "go to" example in these debates, and there is hardly a publication on scientific understanding that doesn't appeal to it.[5] Indeed, Doyle et al. submit that "the frontline of this debate concerns the proper interpretation of the ideal gas law" (2019, p. 345). We resist the temptation to open another front, and we use the ideal gas as our primary example. This is a strategic choice rather than intellectual lethargy. Standard cases serve as touchstones, and arguments need to work in recognised arenas.

The paper is structured as follows. In Sect. 2 we introduce the relevant details of our case study and argue against non-factivism. In Sect. 3 we claim that existing versions of factivism are unsatisfactory because their strategies for accommodating the noetic value of idealised models fail to account for the ways in which an understander (i.e. a scientist) comes to recognise the noetic function played by their idealisation assumptions. In Sect. 4 we provide our positive account of understanding based on the notion of stability and compare it with other styles of scientific reasoning, and philosophical discussions thereof. In Sect. 5 we explore the kinds of understanding that result from such stability with a focus on the contributions made by the accurate and inaccurate aspects of idealised models. Section 6 concludes.

2 The physics of gasses and the flusters of non-factivism

We begin this section by introducing the case of the ideal gas (Sect. 2.1) and then argue against non-factivism (Sect. 2.2).

2.1 The physics of gasses

Gases have been systematically studied since at least the seventeenth century. Investigations focussed on a few macroscopic variables, typically the pressure p, the volume V, and the temperature τ, of a gas, and scientists investigated their relations

[4] This terminology has become customary. See, for instance, Doyle et al., (2019, p. 346) and Dellsén (2016).

[5] Here is selective list of cases in point: De Regt (2017), Elgin (2017), Khalifa (2017), Potochnik (2017), and Sullivan and Khalifa (2019).

under different circumstances. Milestones in this process are Boyle's Law (formulated in 1662), Charles' Law (formulated in 1787), and Gay-Lussac's Law (formulated in 1802). These are empirical regularities, formulated based on extensive laboratory work. They can be combined to give $pV = c\tau$, where c is an empirical constant. We call this the *empirical gas law*: a measured regularity recognising that the product of a gas' pressure and volume is proportional to its temperature. The studies were done under circumstances broadly similar to those of air at room temperature, and so the law holds at low pressures and high temperatures (i.e. temperatures well above the point at which the gas liquifies). Readers will be familiar with this law in its modern form, $pV = nR\tau$, where n is the number of moles in the gas and R is the universal gas constant (itself equal to the product of the Avogadro and Boltzmann constants) and under its modern name, the *ideal gas law*. We do not at this point use it in this form because we want to emphasise that the macroscopic law is *empirical*, a result that emerges from laboratory work and does *not* depend on theoretical assumptions concerning the constitution of matter.

Now enter the realm of theory. In the second half of the nineteenth century, the atomic hypothesis, the claim that matter is composed of particles, gained scientific traction. Physicists, prominently among them Boltzmann, Clausius, and Maxwell, started constructing microphysical models of gases based on the hypothesis. An important model, still used today, construes the atoms of a gas as particles that do not interact with each other, implying that there are no collisions between them. The gas in this model is now known as an *ideal gas* (Tuckerman, 2023, p. 90). In 1857 Clausius showed that the ideal gas law, $pV = nR\tau$, could be derived from the ideal gas (which gives the law its name) in conjunction with mechanical background assumptions (which we discuss in Sect. 4.4).[6] It's now just a small step to identify the empirical constant c with nR, and so the empirical regularity found in experimental studies turns out to be a consequence of the ideal gas model.

Factivists and non-factivists agree that the model provides understanding of the empirical gas law. So, in our parlance, $pV = c\tau$ is the object of understanding; the ideal gas model is the basis of understanding; and the ideal gas is noetic with regards to the empirical law.[7] Unanimity wanes when we ask how the model earns its noetic status. The obvious problem is that the ideal gas model is patently wrong. Molecules of real gases *do* interact; and moreover, their interactions are important to many aspects of the gas' behaviour, for instance its approach to equilibrium. But how can a model that misrepresents such crucial aspects of its target be noetic?

2.2 Non-factivism and continuous gasses

Non-factivists argue that this question is based on the misconception that understanding requires truth. Chirimuuta characterises non-factivism as

[6] The English translation of the paper is (Clausius 2003).

[7] It's crucial for the discussion to be clear on what the object and the basis of understanding are. Doyle et al. offer a vivid account of the confusion that ensues from mixing them up (2019, pp. 352–353).

> the idea that the achievement of scientific understanding is not essentially a matter of finding out the facts about the system and acquiring true beliefs about it (e.g., the discovery of a true explanation), but that it depends on the creation of scientific representations that aid human cognition—especially through the strategies of abstraction and idealisation—involving departures from representational accuracy (Chirimuuta 2023, p. 80).

Some go further and claim that understanding is not merely *compatible* with falsity but essentially depends on it. In this vein, Chirimuuta notes that the "non-factivist assertion is that idealised models offer epistemic benefits, in their conferring of understanding, that are *not matched by more realistic counterparts*" (*ibid*., emph. added).[8] Doyle et al. call this the "parity condition" (2019, p. 346). Non-factivism garnered considerable support in recent debates and has been developed in a variety of ways.[9]

Non-factivists rightly emphasise that literally false models *can* be noetic, and that every tenable account of understanding must accommodate this. However, in their zeal to bring idealised models into the noetic tent, at least some non-factivsts overshoot the mark. Consider de Regt's "criterion of understanding phenomena", the cornerstone of his account of understanding: "A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency" (2017, p. 92). De Regt imposes no factual criteria on the notion of explanation and emphasises that "a general theory of scientific understanding […] should be pluralistic and independent of any specific model of explanation" (*ibid*., 88). The net result is that the only constraint on understanding as regards the relation between the basis and the object of understanding is empirical adequacy.

This is too permissive. Not all empirically adequate models are noetic (even when they satisfy additional criteria like consistency and intelligibility). To see this, let us perform a thought experiment involving the frontline case of the ideal gas. Assume we live in a world where the continuum theory of matter is correct: matter is continuous, and an object evenly fills the space that it occupies. In this world, a gas in a container is a continuous substance spread uniformly over the available space. Further assume that the empirical gas law, $pV = c\tau$ holds in this world.[10] Finally assume that the world's history of physics unfolded in the same way as in the real world, and so the ideal gas model was proposed as a model of the substance in the container some hundred and fifty years ago. Is the ideal gas model noetic in that continuum world? No. The ideal gas completely misconstrues the nature of matter

[8] We note here that the distinction between truth and representational accuracy, as discussed, for example, by Hubert and Malfatti (2023), is not relevant for our current purposes.

[9] See, for instance, de Regt (2017), Doyle et al. (2019), Elgin (2017), and Potochnik (2017).

[10] In this context, pressure and temperature are defined macroscopically and without reference to the microstructure of a gas. This is in line with the history of physics. As we noted in the previous subsection, laws involving pressure and temperature have been formulated since the seventeenth century, long before a microphysical reduction of these quantities was considered.

in the world of our thought experiment, and a model that represents matter as constituted by discrete particles bouncing around like balls cannot provide understanding of matter in a "doughy" or "gunky" world, irrespective of its empirical adequacy and its internal consistency and intelligibility. If a model gets things downright wrong, it can't be noetic, no matter what other good-making features it possesses.[11]

This thought experiment is not a philosophers' fancy; it closely mirrors the history of physics. Continuum and atomistic views of matter were both live options in the development of the study of gases during the mid-19th Century, with the continuum view actually having the upper hand. In this context, Maxwell makes our point explicit when he observes that if reality is "inconsistent" with the basic assumptions of the atomistic kinetic theory, "then our theory, though consistent with itself, is proved to be incapable of explaining the phenomena of gases" (1965, 378). Maxwell didn't distinguish between explaining and understanding (at least not in the way in which current debates do), and so his claim is tantamount to the observation that if gases are not made up of atoms in motion, then the kinetic theory provides no understanding of their observable behaviour (including, obviously, the empirical gas law).

The same line of thought underpins the infamous dispute between Boltzmann on one side and Mach, Ostwald, and Helm on the other.[12] Positivists like Mach were vigorously opposed to an atomic picture of reality. Boltzmann, by contrast, was confident in the existence of atoms and fought tooth and nail for the atomic conception of matter. The point was so important to him that Cercignani's (2006) biography is entitled "Ludwig Boltzmann: the man who trusted atoms". There is logic in Boltzmann's insistence on this point. Boltzmann championed the kinetic theory of gasses, and he was convinced that this theory would provide no understanding of its subject matter unless the world really consisted of atoms. This is precisely the point of the thought experiment, and it makes clear where de Regt's view is too permissive: empirical adequacy (even combined with internal consistency and intelligibility) is too weak to provide understanding if reality is out of sync with the basic assumptions of the theory.

Not all non-factivists are as permissive as de Regt. There are versions of the view that posit additional constraints, such as Elgin's (2017). She submits that to be noetic models must be "felicitous falsehoods", which must be "true enough". But this invites the question: under what conditions do models meet these conditions? We argue (in Sect. 4.5) that a model is true enough exactly if it is stable in the sense that we require. So, her non-factivism becomes indistinguishable from the version of

[11] Note that we are not claiming that *any* discrete model of a continuous phenomenon, or any continuous model of a discrete phenomenon, will get things "downright wrong" and thus fail to be noetic. There are such models that are noetic, and we discuss one in Sect. 4.3: the Lokta-Volterra model. This model represents discrete populations via differentiable functions defined on the real numbers and so is a continuous model of a discrete phenomenon. Yet the model is noetic. What counts as "downright wrong" depends on whether the discrete vs. continuous nature of the target matters to the object of understanding in a way we explicate in Sects. 4 and 5. We are grateful to an anonymous referee for encouraging us to be explicit about this.

[12] For an account of this history, see Cercignani (2006, Chs. 2 and 3).

factivism we propose.[13] More generally, once non-factivists grant that understanding requires more of the model-target relationship than empirical adequacy, they owe us an account of what the required extra ingredient is. The challenge for non-factivists is whether they can specify this without collapsing into our version of factivism.

3 The shortcomings of current versions of factivism

Contemporary factivists agree that scientific understanding is constrained by the facts in ways that go significantly beyond empirical adequacy. But they differ in how they explicate this, and in where and how they embed idealised models into their treatments of understanding. Some factivists develop accounts without addressing the role of idealisations in much detail (e.g. Grimm (2006) and Dellsén (2018)). Some, e.g. Lawler (2021) and Rice (2019, 2021, Chs. 4, 8, 9) argue that the true explanations that underwrite understanding can be "extracted" from idealised scientific models, and so treat those models as conduits to, rather than constituents of, that understanding. Another position falls out of discussions of scientific representation, where Frigg and Nguyen suggest that when properly interpreted, idealisations are not falsehoods (2021). Others, e.g. Strevens (2017), Khalifa (2017), and Sullivan and Khalifa (2019), distinguish between the idealised and non-idealised aspects of scientific models and develop strategies for accommodating the former directly within a factivist framework.

It is this latter group that is our focus in this section where we argue that Strevens' "kairetic account" (Sect. 3.1) and Sullivan and Khalifa's "splitting strategy" (Sect. 3.2) are either untenable or incomplete, at least as they stand. Restricting our discussion in this way is necessary to keep the discussion manageable (discussing the full spectrum of factivist positions would require a book-length treatment), but also theoretically motivated. Whilst we think that both the extraction view and non-literalism are promising, they require supplementation with a more detailed account of how such an extraction or interpretation is to proceed in particular cases. In such an account, the model's idealisations may end up playing a supporting role to more general appeals to aspects of scientific practice including "background knowledge" and "disciplinary practices", or alternatively, they may provide a way of reasoning about those idealisations directly by appealing to the strategies we discuss here. In fact, this latter option is exactly what is suggested by Lawler (2021), who argues that Strevens' kairetic account can be utilised as a manual for extracting true explanations from idealised models.[14] Hence, understanding where the splitting strategy and the kairetic account go wrong provides a useful springboard both for our own positive view, and for the full articulation of more general extraction and interpretation views.

[13] At least as far as model-target relations are concerned. As noted, there can be differences as regards additional requirement such as intelligibility and reflective equilibrium.

[14] Grimm (2006), who, as noted, does not address the question of idealisation in detail, also appeals to the karietic account as a way of dealing with it in a factivist framework.

3.1 The kairetic account

At the heart of Strevens' (2008) treatment of idealisation is the idea that the details of some feature of the target system (e.g. intermolecular collisions) don't make a difference to the object of understanding, and so idealising them (e.g. representing them as not occurring) communicates this non-difference-making in an auspicious way. Following Khalifa we refer to this as "difference-faking" (Khalifa, 2017, p. 173). Assessing Strevens' account of idealisation (and how it relates to his account of understanding (2017)) requires a brief introduction to his kairetic account of explanation. The core idea is that to explain (and so understand) a phenomenon is to accurately represent its difference-makers (and hence the account is factivist by design). This is done in a two-step process.

In the first step, we start with an arbitrarily detailed *veridical model* of the phenomenon, which provides a derivation of the phenomenon from true premises such that the structure of the derivation mirrors the causal chain leading up to the phenomenon (Strevens 2008, pp. 71–83). However, such a model may fail to be explanatory because it may contain "too much" information. Strevens' example is a causal model of the death of Rasputin that specifies the gravitational pull of Mars on the principal actors involved (*ibid*., 88–89). This influence was real, but it was not a difference-maker. The event would have unfolded as it did even if Mars exerted no force on the actors, and thus the model does not explain Rasputin's death.[15]

So, in a second step, the model needs to be subjected to a "kairetic procedure", which eliminates non-difference-making factors. This procedure can be of two kinds. According to the first, we test whether a sentence can be *eliminated* from the model without breaking the causal entailment of the phenomenon; if so, the sentence is eliminated. We then continue to test and eliminate until all that remains is essential for the derivation (*ibid*., 86ff.). According to the second, we test whether a sentence in the model can be *abstracted*, which, roughly speaking, involves replacing it with a more general counterpart (for example, by replacing a specific parameter value with a parameter range), and again testing whether the causal entailment of the phenomena remains (*ibid*., 96ff.).[16] This abstraction procedure is then repeated until no further abstraction is possible without breaking the entailment.[17] The result

[15] An alternative interpretation of the kairetic account is that the overly detailed veridical model is still noetic, since it does represent difference-makers, but it is noetic to a lesser degree, since the explanation and understanding it provides can be *improved* by removing extraneous detail. We are grateful to an anonymous referee for encouraging us to consider this interpretation, but we think it goes against Strevens' own presentation of the account (*ibid*., 117). Regardless, our discussion below applies *mutatis mutandis* to the alternative interpretation and so we put it aside here.

[16] More specifically, the abstraction procedure requires comparing the initial veridical model M' with another model M, whereby "one model M is an abstraction of another model M' just in case (a) all causal influences described by M are also described by M', and (b) M' says at least as much as M, or, a little more formally, every proposition in M is entailed by the propositions in M'" (*ibid*., 97).

[17] We are suppressing some details of Strevens' discussion, including, for example, issues concerning the path dependency of the order in which the kairetic procedure is employed and how to prevent the abstraction procedure from over abstracting (*ibid.*, Chapt. 3). Such detail isn't relevant to our current discussion.

of performing either of these procedures delivers an explanatory model. Scientific understanding, then, is simply the grasping of such a model (Strevens 2013).

The essence of Stevens' account of idealisation (Strevens 2008, Ch. 8) is to compare idealised models with what would result from applying kairetic procedures to their "veridical counterparts", models that represent that same target but without distortions. Like the veridical models introduced above, veridical counterparts may contain too much detail to be explanatory. So, they too need to be subject to the kairetic procedure, delivering explanatory "canonical models", which are both explanatory and veridical.

What makes idealised models explanatory then is their relationship to the canonical models (of their veridical counterparts). Both correctly specify the same difference and non-difference markers, but the former's idealisations function to "[fill] out certain details left unspecified by the canonical explanatory model [where the] details are filled out in a certain way: the relevant parameters are assigned a zero, an infinite, or some other extreme or default value. This is the idealization's way of asserting that the actual details do not matter" (*ibid.* 318). So if a feature is not represented by a canonical model, we can infer it is not a difference-maker, which, in turn, tells us that if a distorted feature figures in an idealised model, the details pertaining to the idealisation are not difference-makers. In our test case, the idealisation that the particles don't interact accurately captures the fact that the details of the particle interactions don't make a difference to the empirical gas law, which follows from the fact that these details are not part of the canonical model. And given the connection between difference-making, explanation, and understanding, this allows for the ideal gas model to be noetic in virtue of accurately communicating what does and does not make a difference to the object of understanding. One important thing to note here, is that the account does not just claim that idealisations are difference-fakers, it also provides a method for identifying them as such: find the idealised model's veridical counterpart; subject the counterpart to the kairetic procedure; and check whether the idealisation is an extremal "filling in" of the resulting canonical model.[18]

However, this account faces a number of issues. First, one might worry again that the kairetic procedures require decomposing models in such a way that the contribution of each sentence (or model aspect) to the result can be tested "one-by-one" by the procedures. Rice's (2019, 2021, Ch. 5) objection that idealised models are holistically distorted and cannot be so decomposed applies: since idealisations play a crucial role in deriving the model results, they will not, in fact, be eliminated or abstracted by application of the procedures, and as a result, will remain in the canonical model. Whilst we think that the objection is onto something important (which we discuss in Sects. 4 and 5), Rice poses his objection in terms of whether

[18] It is worth noting here an alternative interpretation of Strevens' account according to which it provides an explicative *definition* of (non)-difference-making (and thus the conditions under which an idealisation is a difference-faker), and needs to be supplemented with a hands-on method, like the one we offer in this paper, for testing whether any individual idealisation is such. We are grateful to Michael Strevens himself for this suggestion.

idealised models can be decomposed this way, and argues, via a claim about the Maxwell–Boltzmann distribution in the ideal gas law, that they cannot. Even if Rice is correct about this, the kairetic account properly understood does not require eliminating/abstracting idealisations from idealised models themselves; it requires that they be eliminated/abstracted from their veridical counterparts. Rice's argument does not address this directly.

Having said that, a second worry concerns these veridical counterparts, and so appealing to them to block Rice's objection may be a pyrrhic victory. In principle, there is nothing suspicious about the *existence* of such a model, but in practice, we are sceptical as to whether it can do the work the account requires. To illustrate: in principle there is some description of all the particle interactions in an actual gas, and such a description entails the empirical gas law. But it's a stretch to think that this description is something that we have access to, i.e. can write down and solve and then subject to a kairetic procedure. But maybe this is asking for too much. One might say that all we need is a reason to believe that *were* we to apply the procedure to the model, it *would* result in a canonical model that overlaps in the appropriate way with the idealised model. But what would such a reason be? At the very least, more needs to be said about the strategies for probing, controlling, and eliminating/abstracting such counterparts, and we submit that the alternative positive account we offer in this paper better fits with scientific practice.

What about the procedures themselves? They aim to remove detail from a model, via elimination or abstraction. Taking elimination first: to what extent should we think that the corrected version of an idealisation can be eliminated from a veridical counterpart? Not much. Generally speaking, if an idealisation is present in a model, it's there for a reason: it, or something like it, is needed for the derivation of the model's outputs. In thermodynamics we work with quasistatic processes, processes that are infinitely slow and yet change a system's state in a finite time by tracing a continuous line in state space. This is an idealisation (no such processes exist), but we cannot account for the noetic value of a model involving them by first correcting the model by specifying the actual process and then eliminating all mention of processes in the same way in which we eliminate mention of the gravitational pull of Mars in Rasputin's murder; there has to be *some* process taking the initial state to a later state. Or consider a climate model with an idealised ocean of a uniform 50 m depth, which provides understanding of mean global temperature increases and climate sensitivity. We cannot account for the depth idealisation by first specifying the exact topography of the ocean, and then eliminating all mention of ocean depth: the ocean is in the model for a reason, and eliminating (a corrected version of) it precludes the derivation of the model outputs necessary for our understanding of the model's target. These are not cherry-picked examples, they are the rule rather than the exception in scientific modelling. If something like the idealisation were not contributing to the model results, why would it be assumed in the first place? Our objection here mirrors Rice's (*ibid.*): irrespective of whether we are applying the elimination procedure to an idealised model or its corrected counterpart, the concern remains that we are not entitled to assume that idealisations (Rice's point) *or their corrections*

(our point) can be eliminated without destroying the entailment of the relevant model results.

How about the second kind of kairetic procedure, abstraction? Again, working on the veridical counterpart of the idealised model, here the prescription is to test whether model sentences can be abstracted, and the aim is to construct an abstract model which subsumes the original model and still delivers the desired conclusion. For example: our corrected model of a ball breaking a window specifies the particular weight of the ball, but the window would also have broken if the ball had any weight above 1 kg. So an abstract model replaces the precise weight of the ball with the statement "the weight of the ball is greater than 1 kg".

This strategy seems to work if factors are parametrised by a real number (or a tuple thereof) but it's unclear what it amounts to if there is no such parametrisation. This is a problem because many scientific cases are not like the toy example. Our gas is a case in point. The true particle interactions are unknown; but whatever they are, it is far from clear that there is an "abstract form" of them that still delivers the empirical gas law, and that is such that it can be, in Strevens' terms, "fleshed out" (*ibid.* 97) in one way to deliver the idealised model with no interactions, and in another way to deliver the actual interactions.

Such an abstract model would have to have some "interaction feature" that subsumes the interactions used in a large class of models that provide the desired results. Interactions are formally described through interaction potentials, many of which are considered in the physics literature on gases. In a review, physicist Phil Attard describes the situation:

> One may identify several types of intermolecular potentials including the Coulomb interaction due to a net charge on the molecules, dipole and multipole interactions due to permanent nonspherical charge distributions on net neutral molecules, short-range core repulsions ... and long-range dispersion attractions due to induced dipoles arising from correlated electronic fluctuations (2002, p. 153).

Attard's remark that these are *types* of potentials is crucial. Short range repulsions, for instance, include a whole array of very different potentials like hard-sphere, square well, and Lennard–Jones potentials. Attard further notes that of all these "only the Coulomb potential is strictly pairwise additive"; all others are effective potentials for which "it is an approximation to neglect the many-body contributions" (*ibid.*), and so a full list of the potentials that the canonical model would have to subsume needs to include multi-particle potentials as well. To complicate things further, in some situations particle interactions with external sources also play a role in the system's behaviour (*ibid.*, 157).

One way to make the kairetic procedure work would be to find an interaction potential of the form $f(v_1, \ldots, v_k)$, where $v_1, \ldots, v_k$ are relevant physical variables, such that $f(v_1, \ldots, v_k)$ turns into all of these potentials for certain values (or ranges of values) of $v_1, \ldots, v_k$. We have not found such a function in the extensive literature on

the statistical mechanics of gasses.[19] In fact, the physics literature proceeds "inductively" by painstakingly carrying out calculations for one potential after another, which would be a waste of time and effort if there were an abstract $f(v_1, \ldots, v_k)$.

Are we asking for too much? One might argue that the kairetic account doesn't need such a mathematical function. Instead, it just needs an abstract description that says what it is about the actual potential that matters for deriving the ideal gas law.[20] For instance, to understand a certain mechanical problem it may be enough to know that the force function is conservative, and no "superfunction" is need that yields all possible specific force functions through adjusting parameters. Hence, what the kairetic account requires is the identification of an abstract notion like "conservative" that delivers the result while being allowed to "forget" a lot of the detail contained in each of the potentials, and one might argue that such a description need not even be mathematised.

We remain unconvinced that this is workable suggestion. Whilst it is true that the abstraction procedure may move from a specific mathematical potential to a vaguer description in ordinary language, it is crucial for the kairetic account that the result of the procedure, the canonical model, still have the inferential machinery required to derive the object of understanding. In our example, the canonical model needs to have a rich enough inferential structure to continue to entail the empirical gas law. But it is simply unclear how this would be done without something *like* the machinery found in a mathematised potential. This point can be further supported by recalling Cartwright's discussion of abstraction, in which more concrete concepts "fit out" abstract concepts (1999, pp. 39–40). For example, *work* is fitted out via *washing the dishes*, *writing grant proposals*, and *negotiating with university deans*, just as *force* is fitted out via *gravity*, *friction*, and *electrostatic repulsion*. Cartwright's point then is that nothing follows from abstract laws alone, and abstract theories always need to be fleshed out in concrete terms to provide results. Hence, on Cartwright's account there is no abstract model, let alone an ordinary language description, with the same implications as the concrete model we started with.

In sum, Strevens' account of how to determine whether idealisations are difference-fakers relies on the following promises: we can find, or at least reason about, "veridical counterparts" of idealised models; we can make sense of the idea of applying the kairetic procedure(s) to these models; were we to do so, such procedures would result in abstract "canonical models" which don't include the idealised details; and those canonical models would be the sorts of things that entail our objects of understanding. Our point is that none of the above are moves that one

[19] There is a Pickwickian way of constructing such a function: multiply each potential with a parameter and add them all up; the parameters can then be used as "switches" to activate and deactivate certain contributions. This will not do. First, writing down a weighted sum requires knowledge of all summands, and it's unlikely that science will ever get to the point where all possible interaction potentials are explicitly known. Second, proofs would not be forthcoming for such a function because physicists would effectively have to go through it term by term. This then amounts to an inductive method of checking the stability of the object of understanding across model variation, which is at the heart of our positive account developed in the next section, and which defies the spirit of kairetic abstraction.

[20] Thanks to an anonymous referee for encouraging us to be explicit about this.

typically finds in scientific practice. As applied to our example, we cannot disprove the possibility of reasoning from some detailed actual potential $f(v_1, \ldots, v_k)$ or an abstract (possibly formulated in ordinary language) version of it that still entails the empirical gas law, but without a more detailed story about how we should reason thus, the account remains incomplete. Moreover, once it is recognised that an account of scientific understanding based on the idea that idealisations are difference-fakers requires an analysis of *how* one determines whether a given idealisation is a difference-faker (rather than just relying on the bare fact that it is) the onus is on the factivist to provide such an analysis.

So we are not claiming that Strevens' "correct then abstract" strategy is wrong, but the aforementioned considerations suggest it is not helpful in dealing with actual cases as they appear in scientific practice. We propose an alternative route that replaces "correct and abstract" with "vary". In this paper we suggest that scientists compare their idealised models with *differently idealised* models (that are readily found, and which require neither veridical counterparts, nor canonical models). If the object of understanding (as represented in the model) remains *stable* across these "horizontal" comparisons, it's this that demonstrates that the model is noetic.

3.2 The splitting strategy

An obvious way to face the challenge of idealised models is to split such a model into two parts, one idealised (and hence literally false) and one non-idealised (and hence true). With the split in place, the non-idealised part provides understanding in a straightforward factive manner, and the idealised part needs to be dealt with in some way or another. This is what Khalifa calls the "splitting strategy" (2017, p. 173). One way of dealing with idealisations, on this strategy, is then to claim that they serve to highlight an explanatory (and hence noetic) irrelevance: "while explanations [i.e. the true parts of a model] cite difference-makers, idealizations flag *difference-fakers*" (*ibid*., 174). Discussing the ideal gas model, Khalifa sees the non-interaction idealisation as highlighting the fact that particle interactions turn out not to make a difference to the object of understanding, even though they are a *prima facie* plausible potential explanatory factor.

To develop this strategy, Khalifa explicitly appeals to Streven's kairetic account (*ibid*., 174). We have discussed this account in the previous subsection and our worries carry over to the splitting strategy in so far as it builds on it. However, Khalifa hints at an alternative way of articulating the strategy, and this way is further developed in Sullivan and Khalifa (2019). The point of departure here is the so-called virial expansion in statistical mechanics. We discuss this expansion in detail in Sect. 4.4. What matters in the current context is that this expansion shows that the gas law has the general form $pV/c\tau = 1 + \imath$, where the term $\imath$ is a function of the particle interactions. If there are no interactions, the term is zero and we recover the empirical gas law. What is more, this term is "vanishingly small" for systems like dilute gasses at room temperature and normal atmospheric pressure, which justifies neglecting it (*ibid*., 677).

However, contrary to what the "irrelevance flagging" approach might suggest, the virial expansion implies that particle interactions *do* make a difference, and that difference is quantified by ι. So interactions are not completely irrelevant. Sullivan and Khalifa are aware of this and accommodate it by introducing the condition that only the parts of the idealisation that "approximate their de-idealized counterparts" (*ibid*., 676) are noetic, which leads them to contend that "only idealizations' *approximately true* components provide the epistemic goods that figure in understanding" (*ibid*., 677, emphasis added). So, to be noetic, an idealisation must be approximately true, and the idealisation that particle interactions are zero meets this condition.

Sullivan and Khalifa don't elaborate what they mean by "approximately true", but the case at hand may provide a clue. It's not just that ι is small in situations where the ideal gas model is noetic; as we show in Sect. 4.4, it is also the case that it converges to zero if the strength of the interaction tends to zero. This suggests that approximate truth involves the requirement that a realistic model converges to the idealisation when the model becomes less realistic. We agree with this, and we spell out this condition in Sect. 4.2 in terms of limits being regular. This means that in so far as Sullivan and Khalifa's account is developed in terms of approximate truth, their account is in harmony with the position that we develop in the next section.

As noted by Khalifa himself, the splitting strategy has limited applicability and he submits that there are alternative strategies for the factivist. Among those is the "swelling strategy" according to which a model as a whole is noetic to the extent that a scientist accepts it as effective for reaching a particular goal (2017, pp. 175–181). Whether these alternative dialectical strategies are sufficient for factivism remains an open question (we submit without argument that they are not). But regardless, the need to appeal to them at this point raises a much deeper problem. If there is a plurality of approaches, one needs an account of how a factivist might know which approach is appropriate in a given situation to underpin the noetic value of the model in question. But no such account is provided.

This said, Khalifa's pluralism doesn't do any work in our test case. Sullivan and Khalifa are explicit that the ideal gas should be dealt with via the splitting strategy and the flagging of difference-fakers, a claim we agree with if qualified as suggested. The remaining question for them then is: *how* are we to determine whether the idealisations are difference-fakers? Here Sullivan and Khalifa have little to say beyond the idea that the idealised model can be compared with a veridical model. As we will see, this is not the way that scientific practice typically proceeds, and in many scenarios may not be an option because a veridical model may not be available.[21] We submit that an account of understanding should offer a positive methodology to establish whether a model is noetic, and (contra Strevens' account) that this methodology should reflect the techniques applied by scientists to reason about their idealisations.[22]

[21] We emphasise that we're not questioning the *existence* of such model. To carry out a comparison, scientists would have to *possess* such model, and that is rarely, if ever, the case. We clarify this point below.

[22] For readers whose interests lie in epistemology rather than the philosophy of science, another way of thinking about our claim is that we are pointing to understanding's internalist flavour: being able to grasp that an idealisation is a difference-faker at the very least increases (for the arch internalist: is neces-

4 Stable understanding

We now present our account of the noetic value of idealised models (Sect. 4.1), which relies on the idea of exploring a space of models to see whether the object of understanding, or something "close" to it (Sect. 4.2), is stable across model variations. As we will see, our account refines existing discussions of the understanding provided by "universality explanations", and moreover unifies seemingly diverse styles of scientific reasoning including (model-based) robustness analysis, renormalisation group techniques, and the principle of stability (Sect. 4.3). We then explain how it this plays out in our test case (Sect. 4.4) and say why more constrained versions of non-factivism such as Elgin's risk collapsing into our factivist account (Sect. 4.5).

4.1 Stable understanding

Let M be a model of target system T, and let R be the feature of T that we wish to understand. So R is the object of understanding. Let us assume that M entails R.[23] In our example, M is the ideal gas model, T is a real-world gas (roughly at room temperature and standard atmospheric pressure), and R is the empirical gas law. As noted in Sect. 2, T has R and the ideal gas model entails R. The question now is: what condition must M satisfy to be noetic with respect to R?

Above we asserted that variation was the key to understanding. Rather than correcting and then eliminating and/or abstracting features, we "perturb" the details of M. For instance, we can vary the shape of the particles and turn them from non-interacting points into little colliding hard balls; or we can assume that they are balls with a weak long-range interaction; or …. We then check whether the entailment of R is preserved across those perturbations. So, rather than investigating M in isolation, we consider a class of models "around" M which results from varying some features of M while keeping others constant, and then we check whether those models entail R. If the entailment holds for all models in a suitably defined class of models, then M is noetic.

To articulate this idea, we need three notions in place. First, we say that R is *stable* in a class K of models if all models in K entail R. Second, when considering such a class, the qualification "suitably defined" is crucial. To express the qualification, we introduce the notion of a *noetic core* C_M of M. The noetic core is a

Footnote 22 (continued)
sary for), one's understanding, compared to the scenario where the idealisation is, in unbeknownst fact, a difference-faker. *Mutatis mutandis* for the other options Sullivan and Khalifa suggest.

[23] It is customary in the literature on understanding to adopt a linguistic construal of models (the kairetic procedure checks whether *sentences* can be omitted without breaking the *entailment*). We follow this convention as nothing in our account of understanding hangs on how exactly models are understood. Those who prefer an objectual account of models can replace "M entails R" by "M instantiates R" or "M possesses R". For a discussion of different construals of models, see (Frigg 2023). Note that this way of thinking implies that M and T can both instantiate R. We are grateful to Federica Malfatti for encouraging us to be explicit about this.

collection of features of M that scientists propose are indispensable to the model's noetic functioning, and thus these features are held fixed across K.[24] In the case of the ideal gas, the noetic core consists of the posits that (i) gases consist of discrete particles that remain unchanged over time; (ii) particles obey the classical laws of motion; (iii) particles interact through forces which are such that particle collisions are elastic (i.e. preserve energy and momentum); (iv) the temperature of a gas is proportional to the average kinetic energy of the particles; (v) the pressure of a gas is the particles' momentum transfer to the walls of vessel; (vi) the volume of the gas is the volume of the vessel; and (vii) the intermolecular forces are small.[25] The relevant suitably defined class K then consists of all and only those models that have the noetic core C_M. Third, the *perfect model* M_p of T is the model that provides a truthful mirror image of T in all relevant respects in the context in which M is studied. In our case it is a particle model with the true particle interaction. As stressed above, a perfect model is not usually available in practice, but this does not preclude its existence. It is simply what an omniscient being (like Laplace's demon) would formulate when considering T.

Before stating the general conditions for a model to be noetic, a qualification is needed. So far, we assumed that R is a feature of T and that M entails R. That is, we assumed that the model entails the *exact* object of understanding. In some cases, this is too strict because some models in K imply a property R', close, but not identical, to R, and requiring that M imply R exactly will pull the rug from underneath the analysis. So we should only require that M entail some R' which is close to R, rather than R itself. We write $R' \approx R$ to express this closeness. Trivially R is close to itself ($R \approx R$), and so the case in which all models imply R exactly is a special case of the more general case where models imply some R' such that $R' \approx R$. Standards of closeness are determined by the context of investigation. Yet, as we will see in the next subsection, every acceptable closeness relation has to satisfy the stringent condition that R' have a regular limit.

We can now state the conditions for a model to be noetic:

> *Stable Understanding* (SU): Let M be a model of target T, where M has the noetic core C_M. Let R be a feature of T, where R is singled out to be the object of understanding. Let M entail a feature R' such that $R' \approx R$. Furthermore, let K be class of models that consists of all and only those models that have noetic core C_M (hence M is in K). Finally, let M_p be the perfect model of T. Then, model M is noetic with regard to R if, and only if, the following two conditions are satisfied:

[24] Note that at this point, since we are concerned with strategies for identifying the role(s) of idealisations, a model feature is included in a *proposed* noetic core in virtue of scientists deeming it so, i.e. a scientist proposing that the feature is indispensable for the model's noetic entailment of R. In Sect. 5 we discuss the conditions under, and the relevant sense in, which such proposals are successful. We are grateful to an anonymous referee for encouraging us to be explicit about this.

[25] Hill notes that if particles interact with long-range forces, the pair potential must be proportional to r^{-n} with $n > 3$ for the model to entail something close to the empirical gas law (Hill 1960/1986, 323). We return to what "something like" means shortly.

(SU1) *Stability*: Every model in K entails an R' such that $R' \approx R$.

(SU2) *Truth inclusion*: The perfect model M_p is in K.

SU1 requires that there is no model with the noetic core C_M that fails to entail a feature close to R.[26] We note that this feature need not be the same for all models in K; in principle each model can entail a different R'. This raises the question of how SU1 can be established. In a perfect scenario, one could specify K explicitly and go through all models. One way this can be done is when the class K is parametrised, meaning that K results from varying a set of parameters of a model over a certain range. Establishing SU1 then amounts to showing that the implication from the model to the feature R' (for some $R' \approx R$) is stable under perturbations of the parameter values.[27] Another way is to "renormalise" elements of K and show that the implication is preserved, and therefore common to other elements of K.[28] However, in many cases K isn't parametrised and can't be specified explicitly, or doesn't consist of the appropriate kind of mathematical objects to submit to a renormalisation procedure. In such cases, arguments for the conclusion that SU1 holds will then have to be inductive because only a sample of elements from K are available for inspection, and the validity of SU1 is concluded based on the sample. Inductive inferences of this kind are common in science, and they pose no more of problem to SU than to any other area of (the philosophy of) science. SU2 ensures that M belongs to a class that doesn't get things downright wrong. As we have seen in Sect. 2.2, a kinetic model isn't noetic in a continuum world. The requirement that the perfect model be in K ensures that this is the case.

4.2 Restricting admissible closeness relations

Relaxing the requirement that models entail R exactly and requiring only that they entail a feature close to R is a concession to practice, and, as we will see, it is one that is needed to account for the noetic force of the ideal gas.[29] Yet, one may worry that by making room for models in K to imply R' rather than R itself we have opened the floodgates to arbitrariness, trivialising SU.

[26] Note that the stability required by SU1 is stability of an object of understanding across K, which at first blush seem like a different idea that the one suggested by our account's name: "stable *understanding*". As we discuss in Sect. 6, we think there are important connections between the stability of the object across K and the stability of the epistemic achievement, i.e. the understanding, of the user of an idealised model embedded within this class. We are grateful to Federica Malfatti for encouraging us to be explicit about this.

[27] Establishing results of this kind is the aim of a field called *sensitivity analysis*. For a philosophical discussions of sensitivity analysis, see, for instance, Bokulich and Oreskes (2017).

[28] Establishing results of this kind is the aim of *universality reasoning*, championed in the philosophy of science by (Batterman and Rice 2014; Rice 2021) amongst others. We discuss this in more detail in Sect. 4.3.

[29] Gas models are no exception. The generalisation of stability considerations to include features that are somehow close is common in many contexts. For a discussion of such cases in the context of climate modelling, see Harris and Frigg (2023).

While closeness is contextual, there is a general constraint on admissible closeness relations that blocks trivialisation. The formal articulation of this constraint is beyond the scope of this paper; here we give an intuitive characterisation and illustrate the condition with a toy example. The constraint is that R' has a regular limit and that the perfect model must be located somewhere "on the way" to the limit.[30] Models typically depend on certain parameters, and idealisations involve taking these parameters to certain limiting values. For instance, a model of skier moving down slope will depend on friction, and an idealised model sets friction to zero (thereby assuming the motion is frictionless).[31] Let's call this parameter α, and the model that depends on it $M(\alpha)$. R' is a property of $M(\alpha)$, which typically depends on the value of α. In the case of the skier, R' can be, for instance, the time needed to reach the bottom of the slope. A model usually assumes a fixed value for α. In the case of our gas models, α could be the cross section of a molecule, and the ideal gas assumes that $\alpha = 0$. The requirement that the limit of R' be regular means that it reaches the model value "without jumps". This is tantamount to saying that if we perturb our idealised models (by varying their assumptions), "small" perturbations should only result in "small" changes to our object of understanding. If "small" perturbations result in entirely different model behaviours, then the model in question is not noetic, at least on our account.

To illustrate the requirement, consider a toy example.[32] Assume you have a wavy playground slide whose shape is shown in the top right of Fig. 1. You want to understand the slide's length (i.e. the length of the slide is your object of understanding). You construct a model in which the slide is an inclined plane, as seen at the top left of the figure. The model is $M(\alpha)$, where α is a parameter for the "waviness" of the surface and in the model $\alpha = 0$. R' is the length of the model's slope. You immediately realise that in your model R' follows from Pythagoras' theorem and is $\sqrt{a^2 + b^2}$. But you also realise that length of the real slope assumes a value that differs from $\sqrt{a^2 + b^2}$. Our condition now requires that as the waviness of the slope varies, its length varies continuously. We see that this holds: making a straight line just a little bit wavy makes its lengths deviate from $\sqrt{a^2 + b^2}$ by just a little bit, and if you keep increasing the waviness you will at some point hit the perfect model. Hence, the model meets the condition.

Now change the example slightly. Rather than studying a wavy slide you study a staircase. The stairs seem to be just like the slide and so you think you can use the same model. But now you interpret α as the number of steps, and since a straight line can be thought of as stairs with an infinite number of steps, in the model $\alpha = \infty$. As before, R' is the length and is $\sqrt{a^2 + b^2}$. Moving away from the model here means considering a model with a very large, but finite, number of steps. Our condition requires that when passing to such models, the values of R' move away continuously

[30] For a rigorous discussion of regular limits and their philosophical importance see Butterfield (2011); for a discussion of the role they play in an analysis of idealisation see Frigg (2023, Ch. 12).

[31] For a detailed discussion of this example see Nguyen and Frigg (2020).

[32] The example is adapted from Nguyen and Frigg (2020).

from $\sqrt{a^2 + b^2}$. But this fails. The length of a finite staircase is $a + b$, no matter how many (finite) steps it has. So, there is a "jump" in the values of R' when we move away from the model. The limit for the number of steps exists, but it is singular. The model fails to meet the condition, and so we cannot understand the length the stairs with the model.

4.3 SU and related accounts

One way of thinking about our account is that SU1 demands that the object of understanding, or something "close" to it, be "universal" across K, and therefore SU relies on "universality" reasoning (formalised in physics through the "renormalisation group"). In a sense this is correct, but as we will see in this subsection, our account both refines existing discussions of the explanatory and noetic value of such reasoning (Batterman & Rice, 2014; Rice, 2021) and moreover serves to unify other styles of scientific reasoning namely robustness analysis, and the principle of stability (and there may be others).

Originally developed in the context of quantum field theory and statistical mechanics, the renormalisation group (RG) is a mathematical technique to show that a certain behaviour is universal, i.e. stable across a variety of targets that differ with respect to their microscopic details. This style of reasoning and its philosophical significance has been discussed in numerous publications, individually and jointly, by Batterman and Rice.[33] In one of these, they illustrate the method with the example of the lattice gas (Batterman & Rice, 2014, pp. 358–365). To understand how a gas flows around a flat barrier, scientists construct a simple and highly idealised model M. Yet M correctly captures the empirically observed behaviour of such gases, and plausibly, explains it. To do so, RG considers a class of systems that share the high-level properties of locality, conservation, and symmetry, but differ in the details. The class is large enough to contain a realistic model of the target. The analysis then shows that all models in the class exhibit the observed behaviour, which is thereby shown to be universal. The original simplified model can then be "employed to understand the behavior of real fluids" (*ibid*., 364), where the model's explanatory (and therefore noetic) force is grounded in the fact that the RG "guarantees a kind of robustness or stability" throughout the relevant model class (*ibid*.). It is now obvious that RG is an instance of SU. The observed behaviour is the object of understanding, the class of models is the class K, the high-level features that all models in that class have in common form the noetic core C_M, and the successful execution of an RG procedure shows that SU1 and SU2 hold.[34]

[33] See, for instance, Batterman (2009, 2000, 2002), Rice (2018, 2021, 2022), and Batterman and Rice (2014).

[34] There is an interesting question about how our requirement of closeness plays out in cases of the kind Batterman (2002) discusses where limiting behaviours are singular. Space constraints prevent us from discussing this in detail, but we take our view to be compatible with Batterman's. When the object of understanding is a particular phenomenon (as in our cases), Batterman requires that limits be regular for there to be an explanation of the phenomenon solely in terms of features before the limit. This becomes clear, for instance, in his discussion of optics where he insists that aspects of the rainbow cannot be

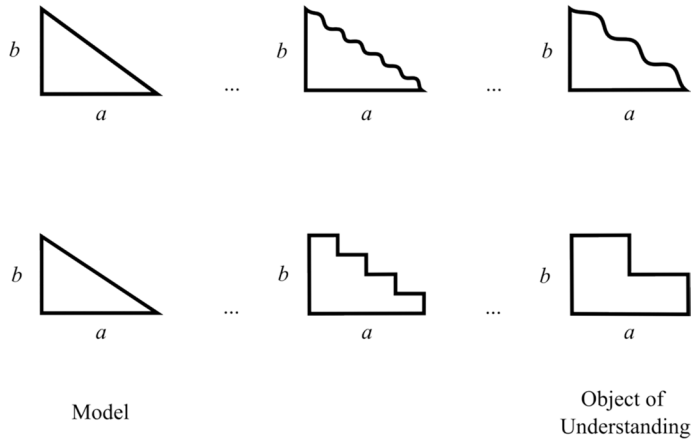

Fig. 1 The playground slide and the staircase

RG is a proper subset of the cases covered by SU. The reason for this is that RG is based on a mathematical formulism that is not present in all cases of SU reasoning. Rice's discussion of RG illuminates this point:

> In applying the renormalization group, one constructs an abstract space of possible systems described mathematically as a space of Hamiltonians [...]. Next, one induces a transformation on this abstract space of systems that has the effect of eliminating various degrees of freedom that are irrelevant to the stable macroscale behaviors of the system. [...] Physicists then look for fixed points under these transformations such that further applications of the transformation yield the same system. When multiple systems all flow to the same fixed points under repeated applications of the renormalization transformation, those systems are said to be in the universality class. (Rice, 2021, pp. 78–79)

We agree with Rice that these techniques are extremely powerful where they apply, but it pays noting that many models that are noetic in the sense of SU are not dealt with in this way because they don't come embellished in the formalism that RG requires, and the relevant model class is not delineated through a class of Hamiltonians. The ideal gas is case in point (we discuss the physics used in that case in the next subsection), as are the Lotka-Volterra model of predator–prey interaction and the climate models used by the authors of the IPCC reports. Models like these are dealt with through robustness analysis (to which we turn shortly), not RG.

Footnote 34 (continued)
understood through wave optics alone because these aspects inhabit "an asymptotic borderland between theories" and therefore a "third explanatory theory is required for this asymptotic domain" (*ibid*., 6) because the "full explanation of what we observe in the rainbow cannot be given without reference to structures that exist only in an asymptotic domain between the wave and ray theories of light" (*ibid*., 21). So the regime at the limit is not understandable solely in terms of what happens before the limit.

Rice is aware of this and notes that his view "requires us to generalize the concept of universality (and universality classes)" (2021, p. 155) and he does so by defining "universality" as the stability of certain behaviour (like our R') across systems that are heterogeneous in their features (like elements in our K, who differ with respect to features not in C_M). "Universality classes" are then characterised as groups of systems with that stable behaviour. He uses this generalisation to develop a "counterfactual account of explanation" (2021, Chaps. 4 and 6) which is then coupled with a (factive) account of understanding (2021, Ch. 8). At this point then, it may be useful to comment on the relationship between SU and Rice's discussion.[35]

It should be noted that SU and the universality account have much in common at a general level. Both recognise that understanding a phenomenon requires exploring a space of systems (our K and Rice's universality classes) that share some features (those in our in C_M) but differ with respect to others (those that vary across systems in the space). Moreover, both recognise the positive noetic value in grasping that features in which the models differ don't make a difference to the object of understanding. However, when it comes to the more fine-grained details of the accounts, we submit that SU goes beyond, or refines, Rice's discussion in at least the following ways.

First, as Rice recognises, when this style of reasoning is employed in non-RG contexts, scientists are relying on "the empirical fact of universality to link scientific models with their target systems", which is different from "the use of mathematical techniques *to explain universality itself*" (2021, p. 156, original emphasis). In other words, one relies on M and T being in the same universality class, rather than explicitly constructing the class and demonstrating that this is the case. But in such contexts, if one aims to understand why the target exhibits R, in terms of the features in C_M being difference-makers to R' and the features not in C_M being-non-difference-makers, one needs a clear account of how to reason about the relevant space of systems, how to determine which features are to be held fixed, how to determine which features are allowed to vary, and how "close" the resulting model-outputs need to be to our original object of understanding for our model to be noetic. Our account is designed to capture how this comes about: start with the idealised model M and identify C_M, vary other features, and check whether R' is stable across the result (we add more nuance to this discussion in Sect. 5.3). In contrast, whilst Rice discusses a nice variety of examples (2021, Ch. 6) he neither attempts to characterise their abstract structure in the way we do with SU, nor does he address the relationship between R and R'. Indeed, in Rice's account there is no distinction between the model-output and the target of a universality explanation, and, *a fortiori*, there is no requirement about how the two relate, let alone that they relate through a regular limit as articulated in Sect. 4.2.

Second, Rice's discussion of universality explanations is coupled with his rejection of what he variously calls "common feature accounts" (2014, pp. 351–357) and the "standard view" of, or "standard approach" (2021, Ch. 1) to, idealisation. Much

[35] Thanks to an anonymous referee for encouraging us to explore this comparison.

like our characterisation of factive accounts of understanding, according to these approaches the explanatory and noetic value of models turns on their accurate representation of the relevant difference-makers to the explanandum or object of understanding. Rice argues that such approaches are inapplicable to universality reasoning, since the models in question cannot be "decomposed" into their accurate and inaccurate parts (Rice 2019; 2021, Ch. 5). As a result, he suggests that universality reasoning provides an *alternative* way of characterising the epistemic value of scientific models, as compared to the standard approach (2019, pp. 196–206). We discuss how we disagree with Rice's claims about decomposition in Sect. 5.2, for now it suffices to note that SU is not committed to a rejection of the standard approach. To see this, first observe that SU2 requires that both the idealised model and the target share the features in C_M (and that this plays a crucial role in the account), and that they differ with respect to features outside of C_M. Then, when coupled with our preferred account of scientific representation, the DEKI account (Frigg & Nguyen, 2021), these points can be mobilised to justify both interpreting the model's C_M features as accurately representing the target's C_M features (like common-feature accounts), and more interestingly, interpreting the features that vary across K as accurately representing features that don't make a difference to the object of understanding (and accurately representing them as not making a difference). For instance, the ideal gas model's point particles *appropriately interpreted* accurately represent that actual gas molecules' extensions don't make a difference to the empirical gas law. As a result, SU shares the standard approach's emphasis on accurate representation.

A third relevant connection between SU and Rice's account emerges from the previous two. Whilst he rejects the requirement of accurate representation on explanatory models he embraces a factive account of understanding (and indeed the content of an explanation). As mentioned at the beginning of Sect. 3, these positions are made consistent by distinguishing between the representational content of scientific models and the content that scientists extract from such models which must then be included in their explanations and understanding (Rice 2021, Ch. 5). Whilst the former might be holistically distorted and generally inaccurate, the latter might be factive. The question then, is how one can extract true (modal) content from idealised models and universality classes. According to SU, which requires that we can distinguish between different model features, it is fairly straightforward to see how we can generate claims like "features in C_M are required for the object of understanding" and "features not in C_M are not", namely by appealing to the features in C_M. By contrast, Rice insists that models are holistically distorted and can therefore not be decomposed into an accurate and an inaccurate part (2021, Ch. 5). Yet, he insists that understanding is factive, a position he refers to as "understanding realism". The core of this position is "that scientific understanding is factive because in order to genuinely understand a natural phenomenon *most* of what is believed/accepted about that phenomenon […] must be true" (2021, p. 251). However, it remains unclear how factive content can be extracted from idealised models and how scientists are supposed to obtain true beliefs from holistically distorted models given that such models said to be holistically distorted (as discussed in the second point). Rice does not address this problem explicitly and instead expresses confidence that it will be resolved

in a "case-by-case approach that allows for a plurality of context-sensitive ways that one's understanding might meet this factive requirement" (*ibid.*). We prefer not to pin our hopes on a case-by-case resolution and insist that an account of understanding requires a structured method to execute such an extraction, and we submit that SU provides such an account.

A fourth connection brings us to our other topic of focus in this subsection, namely the extent to which SU acts as an umbrella for multiple styles of scientific reasoning. When generalising from the RG cases, Rice discusses multiple realisability, unification in Kitcher's sense, and invariance in Woodward's sense as notions that can be captured by a generalised sense of universality (Rice 2021, pp. 180–84). We agree that it is productive to think about these notions as instances of universality or stability. Here we want to subsume two more styles of reasoning, namely robustness analysis and reasoning according the principle of stability as instances of SU. Rice doesn't address the latter, and his treatment of the former has a difference in focus to ours. Specifically, he (i) charges existing philosophical accounts of robustness analysis as being committed to what he calls the "decomposition strategy", which he thinks cannot be applied to models that are holistically distorted (2019, pp. 187–188), and (ii) claims that "robustness analysis is simply not the right tool for solving the problem of inconsistent models" (2021, p. 194). In contrast, since SU recognises that models can be decomposed, at least in the sense that the features in C_M can be distinguished from those not in C_M (more on this in Sect. 5.2), we think robustness analysis is an *instance* of SU, rather than an *alternative* strategy to universality reasoning, as Rice claims.

Moving now from the comparative discussion to showing how these styles of reasoning are instances of SU. Assume an idealised model M of a target T has been constructed and now scientists wonder whether they should take seriously what the model tells them about T. Model-based robustness analysis (MBRA) aims to answer this question by considering a class of models that contains M (Weisberg, 2006). Most models in the class are idealised, although in different ways, and the class is large enough to also contain a truthful representation of the target. A result on which all models agree is a "robust result". The discovery of a robust result is followed by the search for commonalities in the models. If such commonalties are identified, they constitute the "common structure". The final step in the analysis consists in showing that the common structure is responsible for the robust result. An illustrative application of MBRA is the Lotka-Volterra model of predator–prey interaction (Weisberg & Reisman, 2008). It is clear how this maps onto SU: the class of models considered is our K (in the Lokta-Volterra case this is a class including models that represent populations in the aggregate and models that represent individual agents); the robust result is the object of understanding R (e.g. that a biocide favours the prey); the common structure is the noetic core C_M (e.g. that the model populations are negatively coupled); and the fact that all models in the class (including actual

predator–prey populations) entail the object provides understanding in the sense of SU.[36]

This, and other examples of MBRA, are telling because the robust result is often an empirical regularity that was known before the models were investigated, and the models therefore have no novel predictive ambitions. This indicates that the modelling exercise is best thought of as being aimed at understanding. Interestingly, this point is not discussed in the literature on robustness analysis. So the significance of MBRA to understanding becomes clear when we realise that it is an instance of SU.

Let us finally turn to *stability analysis* (SA). Physicists working with mathematical models have repeatedly insisted that structural stability is an important condition on models. Mathematical physicist Robert L. Devaney is explicit: "if the dynamical system in question is not structurally stable, then the small errors and approximations made in the model have a chance of dramatically changing the structure of the real solution to the system. That is, our 'solution' could be radically wrong" (1989, 53). That solutions of a model M must not be wrong in this way, was emphasised by Duhem (2021) and dubbed the *principle of stability* by Fletcher (2020).[37] Stability analysis then aims to establish that a given model is stable under relevant perturbations or variations.[38] SA is an instance of SU: the class K is the result of applying relevant perturbations to M and the elements of the noetic core are the model features that are held fixed in the variation. That the model be stable amounts to requiring that SU1 hold, and that M's result hold of the system itself is SU2.

Hence, RG, MBRA, and SA are all instances of SU. We do not claim that this list is exhaustive, but SU is the backbone of understanding in all of them, and as a result provides a unificatory account of their noetic value.

4.4 SU and the ideal gas

Let us now see how SU's conditions are met in our test case.[39]

In 1857 Clausius published a derivation of the empirical gas law from the ideal gas model.[40] A few years later, Maxwell investigated the equilibrium behaviour of a gas consisting of "perfectly elastic spherical bodies" (Niven, 1965, 378), now known

[36] The conceptual underpinning and different applications of MBRA are discussed in Frigg (2023, Ch. 15) and references therein. For a discussion of the MBRA in climate science, see Harris and Frigg (2023).

[37] The principle of stability posits that "[a]n inference from the statement that a property of a model holds to the statement that the property of phenomena (or some possible world) it represents holds is justified only if all sufficiently similar models also have that property" (Fletcher 2020, 1). For a rigorous mathematical discussion, see Pilyugin (1991); for a discussion of the failure of structural stability see Frigg et al. (2014).

[38] If K is parametrised, this amounts to studying whether the model is stable under perturbations of the parameter values. This is the aim of a subfield called *sensitivity analysis*. For a philosophical discussions of sensitivity analysis, see, for instance, Bokulich and Oreskes (2017).

[39] An anonymous referee encourages us to recognise that Rice offers a detailed discussion of the ideal gas model, showing that the model is holistically distorted (2021). That is not our focus here. We aim to show that (and how) the model is stable in a model class, which is not discussed by Rice.

[40] As mentioned previously, the English translation of the paper is (Clausius 2003).

as the "hard ball" or "hard sphere" model. Maxwell derives the Boyle-Mariotte law from this model (*ibid*., 389), but he doesn't say that it entails the empirical gas law. And this is for good reason: it doesn't. It soon became apparent that in gases with interactions the entailment of the empirical gas law will not be strict. In fact, in 1901 Kamerlingh Onnes (based on previous work by Clausius and van der Waals) formulated what is now known as the *virial expansion*, which states that the general relation between the pressure, volume and temperature of gas is

$$\frac{p}{k\tau} = \rho + B_2(\tau)\rho^2 + B_3(\tau)\rho^3 + \cdots,$$

where $\rho = 1/V$ (Hill, 1960/1986, 261). The coefficients $B_2(\tau)$, $B_3(\tau)$, etc. are called *virial coefficients* and are, as the notation indicates, functions of the gas's temperature. Obviously, if all virial coefficients are zero, this reduces to the empirical gas law.[41] The expansion is telling because $B_2(\tau)$ concerns interactions between pairs of molecules, $B_3(\tau)$ interactions between triplets, etc. (*ibid*.). The virial expansion therefore "contains the idea that the deviations of the ideal-gas law are due to the interactions of the molecules in pairs, triples, etc." (Grossman, 1969, 223). The precise values of these coefficients depend on the specifics of a gas and vary from case-to-case.

The significance of the virial expansion in the current context is that it defines R': a model entails R' iff it entails the virial expansion for some particular values of the coefficients. For this to be a plausible way to understand R', three conditions must be met.

First, the values of the coefficients can in principle be calculated from statistical mechanics, but it's crucial to note that the existence of the coefficients must not be taken for granted.[42] There are gases for which the integrals that define the coefficients diverge (Hill, 1960/1986, p. 262), and hence don't have a virial expansion. K must not contain models of this kind if M is to be noetic.

Second, if the expansion exists, the coefficients must be small. Typically, they are. For instance, for Maxwell's hard sphere model $B_2(\tau)$ takes the value of four times the volume of a sphere that constitutes the gas (Hill, 1960/1986, 268); this is small compared to the volume of the gas, and so the term $B_2(\tau)\rho^2$ will be small. For instance, the Bohr radius of hydrogen is 0.529×10^{-10} m, and so four times the volume of hydrogen atom is of the order of 10^{-30} m^3. The third virial coefficient, $B_3(\tau)$, can be estimated to be of the order of the diameter of sphere to the sixth power (Attard, 2002, p. 200), which is minuscule. Hence, the hard gas ball meets the requirement that the virial coefficients be small.

[41] Doyle et al. (2019), Elgin (2017), Khalifa (2017), and Sullivan and Khalifa (2019) also discuss the virial expansion in this context, but our emphasis is different. Rather than accounting for the noetic value of an idealised model through a comparison with a more veridical model, we focus on the stability of the relationship between the macroscopic quantities involved across perturbations to the model assumptions.

[42] See Attard (2002, Ch. 8), Cowan (2005, Ch. 3), and Hill (1960/1986, Ch. 15) for extensive discussions of how to calculate virial coefficients.

Third, as we discussed above, R' must have a regular limit. In the current context this means that if we consider the limit of R' as the strength of the interaction between particles tends toward zero, that limit must exist and be equal to the behaviour of the idealised model. Assuming that the virial expansion exists (as per the first point), it can be shown mathematically that it has this property (Helrich, 2009, pp. 142–3; Hill, 1960/1986, pp. 261–2): as particle interactions tend to zero, the relationship between a gas' pressure, temperature and volume smoothly tends toward the empirical gas law.

So the third point comes for free, as it were, due to a mathematical result, provided that the first two are met. This raises the question of how the first two points can be established. Unfortunately, there is no "golden bullet". We noted previously that arguments for SU1 will typically be inductive, and gas models are a case in point. Physicists go through the models one by one trying to establish the desired results. Even advanced treatments typically don't cover more than a dozen potentials, which is hardly surprising given the mathematical difficulties that the calculations involve: even for the relatively elementary Lennard–Jones potential no analytic expressions for the virial coefficients are known (Attard, 2002, p. 205). So, any claim that SU1 holds in a relevant class of gas models will be based on a few cases and inductive "in principle" arguments that SU1 will be met in the other cases.

The case of the virial expansion also shows why the requirement that R' have a regular limit is crucial. It requires that if we perturb the ideal gas by just a bit, the behaviour of the perturbed model also moves away from the empirical gas law by just a bit; in fact, the expansion quantifies the deviation and shows how it results from the interactions. If, by contrast, it were the case that any departure from the zero-interaction assumption, no matter how small, would change the volume, temperature and pressure relation completely, then the ideal gas would provide no understanding of that different relation. This would be like trying to understand the staircase with the triangle model. In fact, we are in that situation when particle interactions are so strong that the virial expansion diverges. In such regimes the ideal gas doesn't provide understanding of a gas' manifest behaviour. But this is as it should be: the empirical gas law doesn't hold in those circumstances.

We now turn to SU2, which requires the perfect model to be in K. It's important not to inflate this condition. A scientist who aims to show that a model is noetic must argue that the perfect model is in K, but this does not involve, or even presuppose, that they can explicitly formulate it. The gas is, again, a case in point. There is ample empirical evidence for the conclusion that matter consists of discrete particles that interact with one another through elastic forces. Hence, there is a true force acting on the particles, and this force is the one that features in the perfect model. We don't know this force, but since the noetic core only makes the abstract stipulation that particles interact through elastic forces, we have reason to believe that the perfect model is in K without accessing it.[43]

[43] Returning to our discussion in Sect. 2, we now see why the ideal gas model is not noetic in a continuum world. There, the perfect model is a continuum model. But K only contains models that have the noetic core C_M, which contains the posit that gases consist of discrete particles, and so precludes continuum models. Thus, in a that world, M_p is in not in K, and so SU2 is violated.

4.5 SU and non-factivism

As noted in Sect. 2.2, some non-factivists are less permissive than de Regt and impose model-target constraints beyond empirical adequacy on scientific understanding. A prominent view of this ilk is Elgin's (2017). She holds that to be noetic, models must be "felicitous falsehoods", where "a falsehood or inaccurate nonpropositional representation is felicitous only if it exemplifies features it shares with the phenomena it bears on" (*ibid*., 5). So exemplification requires feature-sharing between model and target, which can be relaxed to allow for situations where "the target does not quite instantiate the features exemplified in the model" but "is not off by much" (*ibid*., 261). In such situations, "the models, although not strictly true of the phenomena they denote, *are true enough* of them" (*ibid*., 261, emphasis added).

But what does it mean for a model to be true enough? Elgin's analysis leaves this largely implicit, but her discussion of the ideal gas contains important clues about what she has in mind.[44] She notes that a perfect description of a gas will have parameters for the description of molecules in it, and "[b]y setting the parameters to zero, it construes the actual size, shape, inelasticity, and mutual attraction of the molecules as negligible", adding immediately that "[s]trictly, of course, in helium the values of those parameters are not zero" and yet "if they are negligible, they can safely be ignored" (*ibid*., 260). What can be safely ignored is then accounted for in terms of the virial expansion (*ibid*., 267). So a gas model is true enough if it has features like consisting of molecules that have size, shape, inelasticity, and mutual attraction in common with the target, and if these are close in the sense that higher-order terms in the virial expansion are negligible. These are the two conditions of SU. They fail in a continuum world, and so in such a world the ideal gas model is not a felicitous falsehood because it is not true enough. In effect, the account draws the line between felicitous and infelicitous falsehoods where SU says that the line should be drawn. Hence, a model is true enough exactly if it is stable in the sense that we require.

Space constraints prevent us from expanding on this, or from discussing other non-factivist positions in detail, but we claim that any non-factivist position faces the challenge: either their view is too permissive (as de Regt's), or they have to restrict admissible falsehoods. In the latter case, we have argued that one such restriction is offered by our SU conditions, but then the resulting position (i.e. Elgin's) collapses into a version of factivism. What other possible restrictions are available, and how they relate to ours, remain open questions.

[44] Elgin also analyses the idea of a felicitous falsehood as "an inaccurate representation whose inaccuracy does not undermine its epistemic function" (*ibid*., 3). But this just pushes the question a level deeper: why do such inaccuracies not undermine their epistemic function, or more positively, how do to they contribute to such a function? We submit our account answers these questions.

5 Two kinds of understanding

The understanding provided by SU is of two kinds, which we call "exclusion understanding" and "inclusion understanding". Intuitively, the idea is that the former concerns features that don't, and the latter that do, matter to our object of understanding, in the sense that the object remains stable on variations on the former, but not the latter. In this section we explore these notions in more detail (Sects. 5.1 and 5.2) and explain how the dialectic between the two offers a methodology of identifying a model's noetic core (Sect. 5.3).

5.1 Exclusion understanding

Exclusion understanding recognises that information about what's excluded can be just as important as information about what's included. If a model feature that seemed to be significant turns out not to be in the model's noetic core, this provides understanding because we then know that R does not depend on it. Exclusion information usually emerges in the construction of the noetic core. As mentioned, models don't wear their cores on their sleeves; and if a model is suspected to be noetic, two things are required: first, a core must be proposed, and second it must be shown that the resulting model class meets SU1 and SU2. Suppose your model involves assumption P (which can but need not be an idealisation). You then find that R is still entailed by models that don't have P. This tells you that R does not depend on P, which is noetically relevant information.

Exclusion understanding appears in our case study. When it was demonstrated that the derivation of the ideal gas law is stable across perturbing "no interactions" to "weak interactions", we increased our exclusion understanding. That is, we found that the empirical gas law does not depend on any specific form of interaction, e.g. molecules being like hard balls, and this boosted our understanding of the empirical law.

At this point we can compare exclusion understanding to other factivist accounts. First, despite their other differences, it should be noted that there is common ground between SU and the accounts offered by Sullivan and Khalifa, Rice, and Strevens, who all agree that there is noetic value in coming to grasp that some features that we might think relevant actually don't make a difference to our object of understanding.

Compared to Sullivan and Khalifa, what our account adds is a story about how one may arrive at such understanding, namely by varying the features and finding that the object of understanding remains stable. Compared to Rice, our account provides a general account of how such understanding is extracted that emphasises the role that accurate representation plays. In particular, SU emphasises that the idealised model and the target share features in C_M despite differing in their other aspects. So (at least when combined with our preferred account of scientific representation) SU delivers the result that the idealised model accurately represents the features in C_M, and accurately represent the other features as non-difference-makers. Exclusion understanding stems from these latter accurate representations. Compared to Strevens, SU offers a different method to identify potential exclusions. Recall that

according to his karietic account exclusions are represented in an idealised model via the fact that they fix a specific value for a ranged property in the corresponding canonical model. But our objection above was that constructing such a model (which requires access to a corrected veridical model, which is then subject to the kairetic procedures, which eliminates or abstracts non-difference-makers P, and results in something which still entail the object of understanding R) is not the sort of thing one finds in scientific practice. Our notion of exclusion understanding also concerns identifying non-difference-makers but it does so in a more liberal manner: they are found by exploring the space of models K and so rather than correcting, then eliminating and abstracting P, and testing whether R is still entailed by what remains, we allow for *replacing* P with an alternative feature Q and testing whether the result is *stable* across the replacement. As a result, we neither require access to the veridical counterpart, nor that an idealised model's noetic core can *by itself* entail R.

5.2 Inclusion understanding

Inclusion understanding comes from grasping what must be included in the noetic core. Assume we don't include an assumption P in the core, and so consider a class K that contains models that don't satisfy this assumption. We then try to establish that SU1 holds but find that at least some of those models don't entail R', and the entailment requires P. This tells us that P is therefore essential to R'. The properties in the noetic core of the ideal gas model are cases in point: the empirical gas law requires that gases consist of discrete particles that don't change through time, that obey classical laws of motion, whose collisions are elastic, and whose interactions are weak. It also requires that temperature be proportional to the particles' average kinetic energy, pressure be the momentum transfer to the vessel's walls, and that the gas's volume is the volume of the vessel. Once any of these assumptions are dropped, or varied, the result fails to entail the empirical gas law. The same applies to Batterman and Rice's example of using the lattice gas model to understand the behaviour of an actual gas: one needs the high-level properties of locality, conservation, and symmetry.

Inclusion understanding should thus be familiar: it corresponds to the understanding provided by identifying difference-makers. As such, it is what is grasped in the demonstration that the object of understanding depends on the features within the model's noetic core. SU goes beyond the splitting strategy and the kairetic account in that it requires neither access to a veridical counterpart, nor the existence of a canonical model, understood as a free-standing model that entails that object by itself. This is crucial to note: in contrast to a canonical model, a noetic core does not need to be a model "in its own right", it is simply what is in common to the models in K that each individually entail R'.

Here it is useful to return to Rice's concern that approaches to idealisation of this sort require an untenable commitment to the claim that models can be "decomposed" in the sense that they can be partitioned according to the contribution of their idealised and non-idealised parts. Rice's concern is explicitly directed at Strevens' approach, since during the "correct and eliminate/abstract" procedure, idealised

model features are (corrected and then) replaced by more abstract features (or are eliminated) that are then part of a corresponding canonical model, while the non-idealised features are left unchanged. The worry then, as Rice puts it, is that: "the parts of the model that are supposed to be accurate representations of relevant features can only make their contributions within the context of the idealized mathematical modeling framework that pervasively distorts them (and many other features)" (2021, p. 136).

It is clear that SU also requires something like this separation: it requires that a model can be "decomposed" into a noetic core, held fixed across K, and a set of features not in the core that vary. Does this mean that SU is vulnerable to the charge that models are "holistically distorted" and one cannot cleanly separate the idealised and non-idealised aspects? No. Rice's objection is to the assumption that the "*contributions* of its accurate parts can be isolated from the contributions of its inaccurate (i.e. idealized or abstracted) parts" (2021, p. 133, emphasis added). We agree that this is untenable. But since SU fully recognises that the non-idealised aspects of an model only "contribute" when *combined* with idealisation assumptions, i.e. that R' may not be entailed by a noetic core alone, but only when combined with some, crucially non-specific, idealisation assumption(s), it does not require that idealised models can be decomposed along "contribution" lines.[45] What is required is the weaker claim that we can *vary* the features that are not part of the core. This is not threatened by Rice's argument.[46] The fact that the features not in C_M are allowed to vary also highlights a crucial ambiguity in Rice's argument against the decomposition strategy (which in turn highlights an important distinction between our account and his, building on our discussion in Sect. 4.3). He states that the problem with what he calls the "standard approach" is that it relies on the idea that "if scientific models are truly decompositional in this way, then the idealizations within our best scientific models should be eliminable in the sense that they could in principle be *removed (or replaced)* without affecting the parts (or contributions) of the model that accurately describe the relevant parts (or features) of the model's target system(s)" (2021, pp. 206–7, emphasis added). And by rejecting this he poses his account as an alternative account. In contrast, whilst SU admits that they cannot be *removed* without affecting those contributions, it positively requires that they can be *replaced* with other (typically still idealised) features, if the model in question is to be noetic and if the features that are so replaced are to contribute to our exclusion understanding.

[45] In contrast, it does seem like the elimination procedure requires this.Whether it is required by the abstraction procedure is an open question and we will not adjudicate between Rice and Strevens here.

[46] Note also that a similar dialectic can be imposed on Batterman and Rice's own example: it doesn't make sense to think of locality, conservation, and symmetry as themselves entailing the relevant object of understanding because they need to be realised by some, by possibly varying, underlying details.

5.3 Identifying the noetic core

By construction, finding out whether a model meets SU requires delineating some class of models with a common core, which vary with respect to some other set of features, and which stably entail an object of understanding, or something close to it. Above, we described this as a two-step process: scientists first identify a set of features they propose as the noetic core and then check whether the result in fact satisfies SU1 and SU2.

In general, we submit, this process involves some trial and error with respect to the noetic core: scientists start with their idealised model, often with some idea about which features might be essential for the model's behaviour. Then they proceed to embark on a creative model building and testing process, building models that differ from their idealised model with respect to features deemed inessential (testing for exclusion understanding), and exploring models which differ with respect to features deemed essential (testing for inclusion understanding). If correct in their initial hypothesis about what did and did not matter for their idealised model to entail their object of understanding (or something close to it), then they'll find that the former models entail R' and the latter don't. In such cases, the noetic core specifies exactly those features that are difference-makers for that object, and the features that vary exactly those that don't. So where SU1 and SU2 are met, the proposed noetic core is the actual noetic core: the set of features that make a difference to the object of understanding, and the features that vary across K are the non-difference-makers.

On the other hand, if their initial proposal was incorrect, this is not a complete failure. Coming to discover that a feature originally deemed inessential is, in fact, required for a model to entail R', or a feature originally deemed essential is, in fact, not so required, is a noetic achievement. In the former case, our original understanding of what didn't matter was mistaken, and coming to realise this increases our inclusion understanding; in the latter case, discovering that what we thought was crucial to the target behaviour was mistaken increases our exclusion understanding.

The way that SU makes room for this iterative process is through the ways in which scientists specify the relevant classes of models. By reasoning about different classes scientists are concerned with different sets of features that are stable. With such new information about whether or not the entailment of some R' such that $R' \approx R$ is stable under some perturbation, scientists adjust what they deem to be the idealised model's core, and thereby consider a new class of models (one that no longer includes models failing to meet assumptions found to be essential, or one that includes additional models meeting assumptions found to be inessential). Of course, whether their proposed core is the actual core of that class of models, and whether it captures the difference-makers of the actual object of understanding, depends respectively on whether SU1 and SU2 are met. It's only by meeting these conditions that a model is properly noetic. But as before, in the ideal scenario, the resulting K is a space of models that agree on all and only the difference-makers, and which differ with respect to all and only difference-fakers, and which includes the actual system. It is therefore a good-making feature of SU that it makes room to account, analytically, for this sort of exploratory process of discovery.

6 Afterthoughts

We have provided a factivist account of understanding that accounts for the role played by idealisations in noetic models in a way that, we submit, avoids the difficulties of extant accounts and squares well with scientific practice. To conclude we want to gesture towards two ways in which SU fits into broader debates.

First, a potential objection. Recall that SU requires that there be a class K, which includes an idealised model M and a number of other models, including a "perfect" model M_p, such that the object of understanding (or something close to it) is stable across K. Thinking of K as a "flat" set of models precludes privileging M as noetic, over any other member of K, including M_p itself. And this is in tension with the observations offered by non-factivists about the *noetic superiority* of idealised models over their more veridical counterparts, as highlighted in Sect. 2. If idealised models are so privileged, and we think they are, how can SU account for this?

The wider debate rightly emphasises other aspects of understanding, for example, the requirement that models be "intelligible" (de Regt, 2017) or allow for an understander to "grasp" the difference-making (and faking) relationships into which our objects of understanding enter (Strevens, 2024). One immediate point is that SU is intended to be compatible with the idea that idealised models possess additional "meta-theoretical" virtues, and it's these that privilege M over M_p (or other models in K). But there is a deeper point here too. Recall from our discussion in Sect. 5 that K is constructed through a creative process of testing to see which features of an idealised model are needed, and which are not, for it to behave the way that it does. In practice, this means that scientists often take M as their reference point, and perturb it to explore a wider space of models. In such cases, this privileges M over other elements of K, since M is the point of departure from which K is constructed through perturbation.[47] In fact, these two ways of arguing for the noetic superiority of M are often related to one another: a reason why M acts as a reference point in the construction of K is typically because it possesses the relevant virtues deemed noetic: simplicity, intelligibility, graspability, and so on.

Second, in addition to connecting SU to the wider debate on scientific understanding, we should also consider how it connects to debates on understanding more broadly. There is a significant question in the epistemology literature concerning whether understanding can be distinguished from knowledge, and one specific battleground of this debate concerns the impact of luck on each of these notions. Those hoping to distinguish understanding from knowledge argue that the former is compatible with epistemic luck in a way that the latter is not.[48] These discussions

[47] It is worth noting that this may be a rational reconstruction rather than a descriptive account of a historical process of modelling. Moreover, we take it to be typical rather than universal. We are grateful to an anonymous referee for emphasising that there may be cases where the idealised model, although still noetic in virtue of meeting SU, is found only after simplifying a more complex collection of information about a target. For a related discussion of how idealised models act to structure spaces of models that turns on the scale of the explanandum or object of understanding see Nguyen et al. (2025).

[48] See, for example, Belkoniene (2022, 2023), Grimm (2006), Khalifa (2013; 2017, Ch. 7), Kvanvig (2003), Pritchard (2010), and Rohwer (2014).

typically proceed via appeals to intuitions about carefully constructed Gettier-like thought experiments involving, for instance, real firefighters surrounded by those in firefighting fancy dress, or accurate books in a library of error-filled tomes. SU provides a way of building bridges between these epistemological discussions and the philosophy of science due to SU1's distinct anti-luck flavour: when met, it ensures that the relationship between an idealised model's noetic core and the object of understanding is stable, and our requirement that R be the regular limit of R' ensures that small changes to a model preclude major changes to the object of understanding. It requires that there's nothing "lucky" about using M, rather than another member of K, to form beliefs about one's target. As such, SU supports the idea that understanding is also subject to an anti-luck constraint. The details are for another time, but combining this observation with the one about how SU recaptures the styles of scientific reasoning discussed above suggests interesting ways of developing epistemological discussions about luck with an eye on scientific practice, rather than artificially constrained thought experiments. In turn this could open a valuable dialogue between philosophers of science and epistemologists of understanding.

Acknowledgements Thanks to Federica Malfatti and Alfredo Vernazzani for organising this special issue and for inviting us to contribute to it. We are grateful to Kareem Khalifa, Michael Strevens, Mike Stuart, and Federica again for comments on earlier drafts. Thanks also to two anonymous referees for this journal for their extensive suggestions for improvement. JN benefited from helpful discussions with audiences in Hamburg, Gothenburg, and Hong Kong; RF from audiences in Miami, Erlangen, Santiago de Chile, and Paris. The paper grew out of discussions in an online reading group on scientific understanding. We would like to thank the members of the group for inspiring discussions.

Declarations

Conflict of interest Both authors declare that they have no conflicts of interest.

References

Attard, P. (2002). *Thermodynamics and Statistical Mechanics: Equilibrium by Entropy Maximisation*. Academic Press.

Batterman, R. W. (2000). Multiple realizability and universality. *The British Journal for the Philosophy of Science, 51*(1), 115–145.

Batterman, R. W. (2002). *The Devil in the Details*. Oxford University Press.

Batterman, R. W. (2009). Idealization and modeling. *Synthese, 169*(3), 427–446.

Batterman, R. W., & Rice, C. (2014). Minimal model explanations. *Philosophy of Science, 81*(3), 349–376.

Belkoniene, M. (2022). *Reassessing Lucky Understanding*. Cambridge Unversity Press.

Belkoniene, M. (2023). Grasping in Understanding. *British Journal for the Philosophy of Science, 74*(3), 603–617.
Bokulich, A., & Oreskes, N. (2017). Models in Geosciences. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science* (pp. 891–911). Springer.
Butterfield, J. (2011). Less is different: Emergence and reduction reconciled. *Foundations of Physics, 41*, 1065–1135.
Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.
Cercignani, C. (2006). *Ludwig Boltzmann: The Man Who Trusted Atoms*. Oxford University Press.
Chirimuuta, M. (2023). Ideal Patterns and Non-Factive Understanding. In K. Khalifa, I. Lawler, & E. Shech (Eds.), *Scientific Understanding and Representation: Modeling in the Physical Sciences* (pp. 78–95). Routledge.
Clausius, R. (2003). On the nature of the motion which we call heat. In *The Kinetic Theory of Gases. An Anthology of Classic Papers with Historical Commentary* (pp. 111–134). Imperial College Press.
Cowan, B. (2005). *Topics in Statistical Mechanics*. Imperial College Press.
de Regt, H. W. (2017). *Understanding Scientific Understanding*. Oxford University Press.
Dellsén, F. (2016). Scientific progress: Knowledge versus understanding. *Studies in History and Philosophy of Science, 56*, 72–83.
Dellsén, F. (2018). Beyond explanation: Understanding as dependency modeling. *British Journal for the Philosophy of Science, 75*, 1261–1286.
Devaney, R. L. (1989). *An Introduction to Chaotic Dynamical Systems* (2nd ed.). Westview Press.
Doyle, Y., Egan, S., Graham, N., & Khalifa, K. (2019). Non-factive understanding: A statement and defense. *Journal for General Philosophy of Science, 50*, 345–365.
Duhem, P. (2021). *The Aim and Structure of Physical Theory*. Princeton University Press.
Elgin, C. Z. (2017). *True enough*. MIT Press.
Fletcher, S. C. (2020). On representational capacities, with an application to general relativity. *Foundations of Physics, 50*, 228–249.
Frigg, R. (2023). *Models and Theories*. Routledge.
Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science, 81*(1), 31–59.
Frigg, R., & Nguyen, J. (2021). Mirrors without warnings. *Synthese, 198*, 2427–2447.
Grimm, S. R. (2006). Is understanding a species of knowledge? *The British Journal for the History of Science, 57*, 515–535.
Grossman, L. M. (1969). *Thermodynamics and Statistical Mechanics*. McGraw-Hill.
Harris, M., & Frigg, R. (2023). Climate Models and Robustness Analysis – Part I: Core Concepts and Premises. In G. Pellegrino & M. Di Paola (Eds.), *Handbook of Philosophy of Climate Change* (pp. 67–88). Springer.
Helrich, C. S. (2009). *Modern Thermodynamics with Statistical Mechanics*. Springer.
Hill, T. L. (1960/1986). *An Introduction to Statistical Thermodynamics*. Dover
Hubert, M., & Malfatti, F. I. (2023). Towards ideal understanding. *Ergo*. https://doi.org/10.3998/ergo.4651
Khalifa, K. (2013). Understanding, grasping, and luck. *Episteme, 64*(1), 1–17.
Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press.
Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.
Lawler, I. (2021). Scientific understanding and felicitous legitimate falsehoods. *Synthese, 198*(7), 6859–6887.
Lawler, I., Khalifa, K., & Shech, E. (Eds.). (2023). *Scientific Understanding and Representation: Modeling in the Physical Sciences*. Routledge.
Nguyen, J., & Frigg, R. (2020). Unlocking Limits. *Argumenta, 6*(1), 31–45.
Nguyen , J., Teh, N., & Shields, P. (2025), Idealisation, Universality, and Explanation, Manuscript
Niven, W. D. (Ed.). (1965). *The Scientific Papers of James Clerk Maxwell*. Dover Publications.
Pilyugin, S. Y. (1991). *Shadowing in dynamical systems*. Springer.
Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.
Pritchard, D. (2010). Knoweldge and Understanding. In D. Pritchard, A. Millar, & A. Haddock (Eds.), *The Nature and Value of Knowledge: Three Investigations* (pp. Part I). Oxford University Press.
Rice, C. (2018). Idealized models, holistic distortions, and universality. *Synthese, 195*(6), 2795–2819.
Rice, C. (2019). Understanding realism. *Synthese, 198*, 4097–4121.

Rice, C. (2021). *Leveraging Distortions. Explanation, Idealization, and Universality in Science*. MIT Press.

Rice, C. (2022). Modeling multiscale patterns: active matter, minimal models, and explanatory autonomy. *Synthese, 200 (Articel Number 432)*.

Rohwer, Y. (2014). Lucky understanding without knowledge. *Synthese, 191*(5), 945–959.

Strevens, M. (2008). *Depth. An Account of Scientific Explanation*. Harvard University Press.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science, 44*(3), 510–515.

Strevens, M. (2017). How idealizations provide understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining Understanding. New Perspectives from Epistemology and Philosophy of Science* (pp. 37–49). Routledge.

Strevens, M. (2024). Grasp and scientific understanding: A recognition account. *Philosophical Studies, 181*, 741–762.

Sullivan, E., & Khalifa, K. (2019). Idealizations and understanding: Much ado about nothing? *Australasian Journal of Philosophy, 97*(4), 673–689.

Tuckerman, M. E. (2023). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press.

Weisberg, M. (2006). Robustness Analysis. *Philosophy of Science, 73*(5), 730–742.

Weisberg, M., & Reisman, K. (2008). The robust Volterra principle. *Philosophy of Science, 75*(1), 106–131.