Micro-level model explanation and counterfactual constraint

Abstract

Relationships of counterfactual dependence have played a major role in recent debates of explanation and understanding. Usually, counterfactual dependencies have been viewed as the explanantia, i.e., the things providing explanation and understanding. Sometimes, however, counterfactual dependencies are actually the targets of explanations. These kinds of explanations are the focus of this paper. I argue that "micro-level model explanations" explain the particular form of a counterfactual dependency by representing the regularity underlying dependency as a necessity in a micro-model.

1 Introduction

Philosophers of science often view counterfactual dependence relations, such as "had x been different, y would have been the case", as central to explanations in science – not least since Woodward (2003). More specifically, counterfactual dependence relations are viewed as part and parcel of the explanantia of explanations, i.e., the things doing the explanatory work. But counterfactual dependencies do not always serve as explanantia; sometimes they are also the explananda, i.e., the targets of explanations. Such explanations are the focus of this paper.

In Woodward's tremendously popular counterfactual account of causal explanation, counterfactual dependencies are supported by generalizations that track causal relations (Woodward 2003, 2018). A generalization tracks a causal relation between variables, e.g. X and Y, if the generalization is invariant under interventions on the 'cause' variable X. For that to be the case, the generalization between X and Y must continue to hold under some range of interventions on X.[1] On Woodward's account we gain understanding of why e.g. Y has a certain value by knowing that certain changes in X *would* lead to certain changes in Y. Woodward also speaks of explanatory generalizations locating the explanandum "in a space of alternative possibilities" and as allowing us to answer so-called "what-if-things-had-been-different" questions, or simply w-questions (Woodward 2003, 191).

A recent trend in the explanation literature has been to not restrict counterfactual dependencies to those supported by causal relations. For example, in Reutlinger's "monist"

[1] Woodward lists a number of conditions that need to be satisfied. See Woodward (2003, 98).

counterfactual theory of explanation is supposed to capture both causal and non-causal explanation (Reutlinger 2016). On Reutlinger's account, an explanation consists of two parts: a statement describing the explanandum phenomenon and an explanans consisting of at least one generalization (and 'auxiliary statements' describing boundary conditions and the like). For the explanans to explain the explanandum, three conditions must be satisfied, one of which is the "dependency condition", according to which the generalizations of the explanans must support a counterfactual. Several other recent accounts of explanation, likewise, identify "non-causal" counterfactual dependencies as central to an explanation's explanans (Bokulich 2011, Saatsi and Pexton 2013, Reutlinger and Saatsi 2018).

I do not doubt that counterfactual dependencies supported by generalizations do important explanatory work in science. But science does more than use such generalizations for the purpose of explanation: science sometimes also makes those self-same generalizations the target of its explanations. One way in which science explains generalizations is by invoking models that postulate micro-entities; these models explain why – given the model assumptions related to the postulated micro-entities – the relationships described by generalizations *must* take the form that they do take. By the same token, such "micro-models" explain the particular form of the counterfactual supported by the relevant generalization. In other words, micro-model explanations *constrain* counterfactual dependencies.

I begin the paper (Section 2) by discussing an example that has figured prominently in many discussions in the explanation literature, namely the ideal gas law.[2] The ideal gas law lends itself easily to a counterfactual analysis – and indeed it has been used by Woodward as an illustration of his counterfactual account. But in physics the ideal gas law is often treated not so much as an explanans but rather as an explanandum explained by the kinetic theory of heat, and later statistical mechanics. Woodward has something to say about this, albeit not much positive.

In Section 3 I introduce my account of *micro-level model explanations* by showing how the kinetic theory of heat explains the ideal gas law. I demonstrate the wider applicability of my account with further examples and demarcate the account from the DN model of explanation and the mechanistic account of explanation, with which it shares some superficial similarities. In Section 4 I further spell out the notion of *representation of generalizations* (or *regularities*, as I will prefer to call them) *as necessities*, which is central to my account. I will also address several worries about the notion. In Section 5 I discuss the problem of explanatory demarcation, which arises from the idealizations that model explanations make. Section 6 concludes the paper.

2 Explaining the ideal gas law

The ideal gas law (IGL) combines the empirically discovered Boyle's law ($P \propto \frac{1}{V}$) and Gay-Lussac's law ($P \propto T$) and has the following form: $PV = nRT$, where P=pressure, V=the volume of the gas container, T=temperature, R=the ideal gas constant, and n=the amount of substance of gas in

[2] See e.g. Friedman (1974), Salmon (1984), de Regt and Dieks (2005), Elgin (2007), Strevens (2008), Doyle et al. (2019), Rice (2019), Sullivan and Khalifa (2019), just to mention a few.

moles. IGL is briefly discussed by Woodward in his book (Woodward 2003). For example, Woodward at one point in chapter 5 states that, on the basis of IGL, one can explain an increase in the temperature of a gas by an increase in the pressure of the gas, because IGL allows one to answer a range of w-questions (p. 221).

In the same chapter, Woodward also comments on the theory that is widely regarded qs explaining IGL, namely statistical mechanics:

> Statistical mechanics does not explain in virtue of doing something different in kind from this [what IGL does], but instead simply provides information that allows us to answer what, in some respects, is a wider, more detailed range of w-questions; hence the sense that in some respects, it provides deeper explanations. (For more on this subject and the "in some respects" qualification, see section 5.12.). (Woodward 2003, 223)

Although Woodward does not spell out what it might mean with statistical mechanics providing "deeper" explanations than IGL, in his view explanations are deeper (than other explanations) if they provide answers to a wider range of w-questions (see Woodward 2003, chapter 5). And explanations provide answers to a wider range of w-questions, in turn, if the relationships in question are invariant under a wider range of interventions (than other relationships might be). The van der Waals equation $\left[P + \frac{a}{V^2}\right][V - b] = RT$, for example, is invariant under a wider range of changes than IGL: it incorporates a parameter for the diameter of the relevant gas molecules (b) and a parameter for the attractive forces between the molecules (a) and is accurate for higher degrees of pressure and temperature than IGL. Woodward therefore considers the van der Waals equation to provide a deeper explanation than IGL (Woodward 2003, 260).

Of course, the van der Waals equation is not statistical mechanics, but if Woodward is right, then what is true of the van der Waals equation should also be true of statistical mechanics (and perhaps even to a greater degree). Yet, in the aforementioned quote, Woodward also flags a qualification regarding statistical mechanics providing a deeper explanation than IGL. This qualification he spells out in terms of what he calls the (hypothetical) "microscopic strategy", which would consist in trying to explain changes in e.g. the value of the pressure variable P upon a change in the volume variable V by considering the initial and final positions of the molecules in the container and the sum of their individual transfer of energy (Woodward 2003, 231-232). The microscopic strategy would try to explain the new value of P by "aggregating the energy and momentum transferred by each molecule to the walls of the container".

Woodward deems the microscopic strategy unsatisfactory because it fails to answer the relevant w-questions: it doesn't tell us determinately what value P would have had, had the initial microstate been different. That is so because any macro state (such as a particular value of P) is compatible with basically any initial state of molecules and its evolution (with different molecule trajectories). Woodward points out that the micro strategy would be "impossible" to carry out, but concludes that even if it were to be attempted, the microscopic strategy "would fail to provide the explanation of the macroscopic behavior of the gas we are looking for", because it "omits information that is crucial to an explanation of pressure" (namely counterfactual information)

(Woodward 2003, 232). Since IGL, in contrast, does identify situations under which things would have been different, it does a better job than the microscopic strategy at explaining the change in pressure.

From all of this, one might get the sense that "going micro" does not provide much understanding over and above the understanding one already possesses by virtue of IGL. This, however, would clearly be at odds with how scientists usually view the situation (e.g. Holton and Brush 2001). But what is this "extra" understanding that micro-models such as the kinetic theory of heat provide? In what follows, I will argue that this "extra" consists of micro-models providing explanations of why regularities obtain in the first place and, accordingly, why they support certain counterfactuals, and not others.

3 Micro-level model explanation and counterfactual constraint

Let a "micro-level model" be a model that postulates unobserved or even unobservable entities for the purpose of explaining empirical regularities and the counterfactuals supported by them. As I see it, there are three components of micro-level model explanations:

1. Explanandum: an empirically discovered **regularity** between macro variables (supporting *counterfactuals*).
2. Explanans: a **micro-level model** postulating entities (and their activities) allows for the derivation of relationships (one might call them model-*internal* relations) which **represent** the explanandum regularities as **physical necessities.** By doing so, the micro-level model **constrains** the form of the counterfactual associated with the explanandum regularity.
3. The model thus provides **conditional how-necessarily understanding**: it tells us why, given the assumptions of the micro-level model, the explanandum regularity has the form that it does have *and not some other form,* and why, accordingly, certain external counterfactuals hold, and not others. Micro-level model explanations, one might say, answer *why-would*-things-*have-been-different questions*, or simply ***ww*-questions** (pronounce: "quadruple-u questions"): they explain *why* a change in x would have resulted in a change in y, had x changed.

Let us unpack these three components on the basis of an example. For convenience's sake, let us use the explanation provided by the kinetic theory of heat (KT) of IGL.[3]

The explanandum in our example is IGL: as already noted, IGL relates macro variables of pressure, temperature, and volume of a gas and supports counterfactuals such as: "had the temperature changed (and the volume stayed fixed), the pressure would have changed" or "had the volume changed (and the temperature stayed fixed), the pressure would have changed". IGL is an *empirical* "law", i.e., it was discovered on the basis of experiments (by Boyle, Gay-Lussac, and others), and as such, it is entirely contingent: in a different possible world, it would have taken a

[3] KT is called a theory for historical reasons. Given its idealizations (e.g. it ignores the extension of molecules and their interactions / collisions) it is perhaps more appropriately described as a model.

different form, such as $P = nRTV$. In such a world, when the volume of a gas is decreased (and T held fixed), the pressure of the gas would decrease (rather than increase, as in IGL). Let me refer to this hypothetical law as the *other worldly gas law*, or just OWGL. But why are pressure, temperature, and volume in our world related by IGL, rather than by OWGL?

This is where KT comes in: it postulates (unobserved) entities, namely moving molecules, which allow the representation of IGL as a necessity. This happens roughly as follows.[4] First of all, let *l* be the distance between two opposite container walls and v_x = the molecule velocity along the x-axis. Then for a molecule on a 'round trip' between opposite sides of the container wall, the number of collisions of a molecule with the container wall per second is $v_x/2l$. The total momentum change per second for all molecules (*N*) in the container with average speed $\overline{v_x}$ is $N\frac{m\overline{v_x^2}}{l}$. By Newton's second law, the change of momentum is equivalent to the average net force exerted by the molecules on a container wall per second: F = $N\frac{m\overline{v_x^2}}{l}$. Since pressure is defined as the force exerted perpendicularly per area (here: l^2), we can divide both sides of the equation by l^2 , which gives us $P = N\frac{m\overline{v_x^2}}{l^3}$. But l^3 is just the volume of a cubical container, so we can write $P = \frac{Nm\overline{v_x^2}}{V}$. Assuming that there is no preferred direction of a molecule's path in the three spatial dimensions, $\overline{v_x^2} = \frac{1}{3}\overline{v^2}$, and $PV = \frac{1}{3}Nm\overline{v^2}$. Since the *total* kinetic energy of translation KE_{trans} of all molecules is $N \times \frac{1}{2}m\overline{v^2}$, this gives us $PV = \frac{2}{3}KE_{trans}$, also known as the *gas pressure equation*. Finally, under the assumption that molecular speed (v^2) depends only on temperature (to be proven separately), it follows that the gas pressure is proportional to the volume of the gas when T is constant (in agreement with Boyle's law and IGL).

What have we gained with KT and the derivation of the gas pressure equation in relation to what we knew by IGL already? First of all, we have derived an expression based on model-internal notions, such as *molecules*, *molecule motion*, *translational energy*, *molecule-container wall collisions*, which holds necessarily conditional on the truth of the model assumptions. Moreover, we (explicitly and implicitly) related this expression to IGL, i.e. we related the macro variables of gas pressure and volume (that figure in IGL) to consequences of KT's postulate of highly idealized moving molecules, and in particular, the translational kinetic energy of molecules. During the derivation we also assumed that the pressure of a gas corresponds to the average net force exerted by the gas molecules on the container walls. Thus, overall, we represented IGL as a necessity (on the basis of the model notions). We can therefore see why IGL holds, rather than some other relationship (such as OWGL), and why certain counterfactuals are supported *and not others*. With KT, we are now in a position to answer *ww*-questions such as "why would P have risen, had the volume been reduced (and the temperature held fixed)?" The answer KT gives is: if the volume had been reduced, the frequency of collisions of molecules with the container wall was *bound to* increase, because less space for the same number of molecules in motion (whilst keeping their speed fixed) *must* result in more molecule-wall collisions.

[4] E.g. See e.g. Holton and Brush (2001) for more details.

The understanding occasioned by micro-model explanations is of a certain type. Traditionally, scientific explanations were thought to give us "how-actually" understanding (Hempel 1965, Craver 2007). More recently, there has been much talk about models providing 'how-possibly' understanding (Grüne-Yanoff 2013, Rohwer and Rice 2013, Reutlinger et al. 2018). Yet the micro-model explanations I'm concerned with here seem to give us something else, namely 'how-necessarily' understanding: they tell us why a regularity *must* obtain (conditional on the model assumptions) and why, consequently, only certain counterfactuals hold (and not others).

3.1 Further examples

In order to demonstrate the applicability of my account of micro-level model explanation, I want to briefly discuss three other examples (in this order): the Bohr model of the atom and the explanation of the spectral lines of hydrogen, Dalton's atomism and the explanation of the laws of constant proportions, and Mendel's explanation of hybridization experiments.

The Bohr model of the atom (proposed in 1913) was devised to explain the empirically discovered spectral line patterns of hydrogen, which were summarized in the so-called Rydberg formula $\frac{1}{\lambda} = R(\frac{1}{{n_1}^2} - \frac{1}{{n_2}^2})$, which describes the spectral emission and absorption lines of hydrogen.[5] This regularity supports certain counterfactuals, such as "had n_2 been changed, the frequency $\frac{1}{\lambda}$ of emitted / absorbed light would have changed".

In order to explain the phenomena summarized by the Rydberg formula, Bohr postulated that, inside the atom, electrons revolve around the nucleus on stable orbits with fixed energy levels. Whenever electrons would "jump" from one orbit to another, the atom would emit or absorb light (depending on whether the "jump" was from a higher to a lower energy level or vice versa). On the basis of the Coulomb's law and several other assumptions,[6] Bohr was able to derive a model-based representation of the Rydberg formula. The constant R could now be specified in terms of the concepts of the model, such as electron mass, charge, the Coulomb constant, and π (for circular orbits). The quantum numbers n_1 and n_2, which in the Rydberg formula had no deeper physical meaning, could now be interpreted as the "order" of electron orbits (starting from orbit closed to the nucleus).

Crucially, the Bohr model explains the Rydberg formula by *representing* it and the counterfactual dependencies supported by it as *necessities*: since it's not possible in the Bohr model for electrons to occupy positions outside of the fixed stationary orbits, the model explains why the measured line spectrum of hydrogen *must be* discrete rather than continuous.

To put the same point slightly differently: the Bohr model helps us to answer *ww*-questions such as "*why would* the frequency have changed non-continuously, had the energy state of the electron in the hydrogen changed". The answer of the Bohr model is: because electrons traverse the nucleus only on discrete, stable orbits. There is no continuous emission or absorption of energy

[5] In the Rydberg formula, n_1 determines the kind of spectral line series, e.g. the Balmer series for $n_1=2$.
[6] E.g. Bohr assumed – quite boldly – that the frequency of emitted light is equal to the average of the frequencies of the electron on its orbit before and after the "jump".

in the hydrogen atom because electrons *cannot possibly* occupy any trajectories outside their stable orbits in the Bohr model.

Dalton's atomism (proposed in 1808) explains the laws of definite proportions and the law of multiple proportions. The former law says that chemical elements always combine in the same ratios. E.g. oxygen always makes up 8/9 and hydrogen 1/9 of the mass of water. The latter law says that whenever two elements form more than one compound, the ratio of those compounds will be multiples of each other, as carbon monoxide and -dioxide where a certain amount of carbon combines with exactly twice as much oxygen in the second compound as it does in the first. A counterfactual associated with these laws might be e.g. "had the amount of carbon been x' (rather than x), the amount of oxygen would have been y' (rather than y)".

Dalton's atomism explains the *form* of the laws and their associated counterfactuals, i.e., it explains not only the fact *that* we observe what is described by the two laws, but also why we do *not* observe any intermediate ratios. E.g. we wouldn't observe that one unit of carbon combines with 1 ½ as much oxygen (rather than twice as much) as it does in CO. Such a combination is ruled out by the assumption of indivisible atoms: it is for this reason that elements combine with each other only in integer multiple proportions. In sum, we can say that Dalton's atomism explains *ww*-questions such as "*why would* the amount of carbon have doubled in a sample of carbon monoxide, had the amount of oxygen doubled (rather than tripled)?" The answer the model gives is that, since atoms are indivisible, there *could not have been* combinations of elements with different proportions.

Let us consider a final example. In hybridization experiments with pea plants, Mendel made some ground-breaking discoveries in pea crossing experiments in the mid-19th century. When he crossed purebred (recessive) white and (dominant) purple flower plants and self-fertilized their offspring, the second filial generation (F2) would consist of a 3:1 ratio of purple vs. white flowered plants. A counterfactual associated with that law might be: "had this cross of flowers consisted of *only* pure-bread dominant or recessive, a 4:0 ratio of either purple or white flowers would have resulted".

Mendel explained the law and the associated counterfactuals by invoking indivisible unitary genetic 'factors' (as he called them) that come in pairs for each organism (alleles). For each hybridized pair of plants there are exactly four possible combinations of alleles (two per plant). Together with the principle of dominance, according to which traits of the dominant genetic 'factors' are always expressed phenotypically, the self-fertilization of purebred recessive and dominant plants from F1 *must* result in a 3:1 ratio in F2. For comparison, a 2:2 ratio in F2 would be impossible on the model.[7] Once again, also Mendel's model answers *ww*-questions such as "*why would* the proportion of red and white flowers in F2 have been 3:1 (rather than e.g. 3.5:2.3), had one crossed red and white plants in F1?". The answer in the model is: because the relevant traits are

[7] The ratios can also be expressed as probability distributions (e.g. 75% vs. 25%). The point remains that Mendel's model represents such distributions as necessities. See also Section 4.3.

inherited discretely on dominant and recessive alleles, making it impossible to obtain ratios that are combinatorially incompatible with a set of 2 x 2 alleles.

I take these three examples to be sufficient to motivate my account.

3.2 MLM-E vs. the DN model and mechanisms

My account of micro-level model explanation, or MLM-E for short, contains elements that superficially resemble other accounts of explanation. In particular, the derivation of model internal relations and the postulation of entities and activities should remind one of the deductive-nomological (DN) model of explanation, and mechanistic accounts of explanation, respectively. I therefore want to use this section to briefly demarcate my MLM-E account from these other accounts.

On the DN model *events* are explained by deriving them from a law of nature and the appropriate boundary conditions. For example, an apple falling from a tree is explained by the apple following Newton's law of gravitation and the appropriate boundary conditions. An obvious similarity between MLM-E and the DN model of explanation is that in MLM-E, too, a logical deduction is involved in the explanation of the explanandum. However this is where the similarities already end. First, on the DN model of explanation the prime target of explanations are events rather than generalizations (Hempel and Oppenheim 1948). Second, as Woodward (2003) as pointed out most forcefully, the DN model requires a universality of laws that is unrealistic in most of the so-called 'special' sciences (and arguably even in the most 'exact' sciences, see e.g. (Cartwright 1983, Reutlinger et al. 2017)): without universality there can be no deduction of the event to be explained. On the MLM-E account, in contrast, there is no requirement of universality regarding the relations in the explanans. Third, on the MLM-E account, the postulation of micro-entities is essential for explanation: the derivation of model-internal relations is based on and motivated by the model postulating micro-entities. There is not anything like that on the DN-model.

On mechanistic accounts of explanation, the explanandum phenomenon (which can be a regularity) is explained by reference to a mechanism that "produce[s], underlie[s], or maintain[s] the phenomenon of interest" (Craver and Tabery 2019). Mechanisms have been defined differently by different authors, but mechanisms are often described as consisting of entities, or parts, and their activities. Obviously, MLM-E explanations also invoke entities and (sometimes) their activities to explain the explanandum. In that sense, MLM-E is superficially similar to mechanistic accounts. There are however at least two crucial aspects in which MLM-E diverges from some of the fundamental assumptions in the mechanism literature.

First, in mechanistic explanations there is no necessity involved: mechanisms explain by showing how the explanandum phenomenon was *actually* brought about. On MLM-E, in contrast, the representation of empirical regularities as necessities is essential for the understanding obtained with micro-models. Second, mechanistic explanations require a realist commitment to the mechanism entities: if the entities in the mechanism are not real or do not exist, then the explanandum phenomenon cannot be caused or produced by the mechanism and – by the lights of

the mechanists – thus not explained. MLM-E, in contrast, is more liberal: often the entities postulated by micro-models are knowingly unrealistic, as in all of the examples discussed above (contra KT, there are intermolecular collisions; contra the Bohr model, there are no electron orbits in atoms; contra Dalton, atoms are surely divisible; genetic inheritance is usually much more complicated than envisaged by Mendel). But if the entities postulated by the model do not exist in the form they are postulated, then the notion that the explanandum phenomenon is explained by entities that *produce* or *cause* them becomes dubious.[8] There is more to say about the issue of idealization and realism; I will return to it in Section 5.

In sum, the MLM-E account is substantially different from other superficially similar accounts of explanation. I should clarify though: I'm not in principle opposed to extending the label of "mechanistic explanation" to MLM-Es. So long as the basic elements of the features identified by my account are respected, this would then merely be a terminological matter.

I will use the remainder of the paper to further clarify my account.

4 Micro-level model explanations and necessity

According to MLM-E, necessity plays a crucial role in explaining regularities and their associated counterfactuals. In what follows, I will further elucidate this role.

4.1 Conditional necessities

Let us first of all stress once more that the necessities of micro-level model explanations are *conditional* necessities, namely conditional on the truth of the model postulates. This is quite different from another recent modal account of explanation, namely the one by Marc Lange's explanation by constraint (Lange 2017). This kind of explanation works "by identifying certain constraints to which the world must conform" (Lange 2017, xvi). For example, mathematical facts constrain the way in which strawberries can be divided evenly among a group of children and conservation laws constrain the kinds of physical forces that are possible.

For Lange, explanatory constraints either explain by virtue of mathematical or physical necessity. For example, it is (presumably) a physical necessity that forces conserve energy. MLM-E, in contrast, explain by *representing* the relevant regularities as necessities. That is, the necessities that figure in MLM-E are not necessarily *actual* physical necessities; first and foremost, they are necessities by virtue of being derived from the model's postulates.

4.2 Necessities and contingencies

According to MLM-E, we understand why a certain regularity and its associated counterfactual holds, because we, in a sense, "replace" the contingency of the regularity with the necessity that springs from the model. One may object, however, that on certain accounts regularities, or "the laws of nature" are necessities themselves, conceived as relations between universals (e.g.

[8] Can one not provide an instrumentalist interpretation of mechanistic talk in science? Colombo et al. (2015) have argued that one can. However even they do not question the idea that mechanisms *produce* their explanandum phenomena.

Armstrong 1983). On such accounts of laws, it would seem, my account of explanation would be unworkable, as we would have necessities explaining necessities (rather than contingencies).

I have three comments about this objection. First, the nomic necessity view of lawhood does not sit particularly well with the fact that many laws of nature are not exceptionless universal generalizations of the form "all Fs are Gs" (as supposed by the view) but rather ceteris paribus laws that hold only under certain conditions and only within certain ranges (Reutlinger et al. 2017). For example, as we mentioned already, the ideal gas law is correct only up until a certain temperature and pressure (above which the van der Waals equation is more accurate). So either the ideal gas law is not a law, or the nomic necessity view of lawhood is too strict for describing the regularities that often are the target of MLM-E.

Second, many have objected that the notion of "non-logical or contingent" necessities seems to be rather mysterious (and perhaps self-contradictory). Third, on the nomic necessitation view of lawhood, the necessitation relation is still (metaphysically) *contingent*. That is, on such accounts it would still be contingent that, in our world, a reduction in gas volume necessitates an increase in gas pressure. Hence, even on such accounts, MLM-Es could still be appealed to in order to explain why the relation has the form that it does have in our world.

In sum, I think the nomic necessitation view of laws of nature is either so problematic that the laws of nature better not be viewed as necessities, or the nomic necessitation view actually entails that laws of nature are contingent (despite the necessitation relation) anyway. In either case, the core idea of MLM-E remains intact.

4.3 Necessities and events

Wesley Salmon once raised an objection against the "modal conception" of explanation, i.e., the conception of explanation that "says that a good explanation shows that what did happen had to happen" (Salmon 1998, 321).[9] This objection is in principle also relevant to the MLM-E account.

Salmon's complaint was that the modal account fails to capture the explanation of events because they have only a certain probability of occurring and *need not* occur. For example, quantum mechanics does not predict (or explain) any particular measurement of a quantum system; it only predicts certain probabilities of measurement outcomes. Likewise, Mendelian genetics does not predict (or explain) any particular color of peas, but only certain probability distributions. Thus, modal accounts that suppose that *events* are explained by reference to a physically necessary law indeed don't look very plausible in the face of such examples.

Is this also a problem for MLM-E? It seems not: MLM-E's target of explanation are regularities, not events. However Salmon insists any account of explanation must address the

[9] Salmon conceived of the modal conception in terms of a relation between a physically necessary law of nature and an event-to-be-explained, which made it sound quite akin to the DN model. He stressed, though, that in contrast to the DN model, on the modal conception the explanandum is not explained by deriving it from a law, but rather by showing that the explanandum event is "physically necessary relative to the explanatory facts" (Salmon 1984, 111). See also Lange (2017) for a modern modal account of explanation.

explanation of events, because "even if we grant the point about theoretical science, one can hardly doubt that applied science often tries to explain individual occurrences or limited sets of occurrences" (Salmon 1998, 323). Salmon probably does not mean "applied science" in the contemporary sense of science applied to some more practical, technological problem. Instead he probably just meant "science applied to specific empirical problems". Be that as it may, science (whether more theoretical or more "applied" in this sense) often *does* explain relations rather than events. When that is the case, it would seem wrong-headed to demand from a philosophical account of explanation to provide an explanation of events. Let us re-consider Mendel's model of inheritance (which, incidentally, is also Salmon's example).

Mendel's model predicts that the ratio or probability distribution of flower color in the aforementioned example will be 3:1. Thus, any given flower in F2 has a 1 in 4 probability of being white. This distribution is indeed necessitated by Mendel's model: provided the 'experiments' are conducted conscientiously, there cannot be any other distribution in F2. The model does not say anything about what the color of any *particular* plant must be, or even what a limited number of breeding experiments will yield. We may of course give further causal explanations of why a certain color was expressed by the genes in any particular plant (maybe contrary to what was expected) or why the actually observed ratio diverged slightly from the predicted probability distribution (e.g. 2.99:1.11 rather than 3:1).[10] However we shouldn't dismiss an account of explanation on the basis of not accommodating such piecemeal causal explanations. *Particularly* not when the scientific model in question doesn't do so either.

The analogous considerations – in principle – apply to the example of quantum mechanics: quantum mechanics predicts probabilities of *possible* measurement outcomes with necessity. I would be prepared to argue that it is (also) by virtue of necessitating these probabilities that quantum mechanics possesses explanatory power. Given the measurement problem, however, this can't of course be the whole story. Then again, quantum mechanics is a difficult case for *any* account of scientific explanation (e.g. Salmon 1998, 325). Even on Woodward's more liberal account of causation, which is not committed to any kind of transmission between cause and effect, quantum mechanics is not accommodated (de Regt 2004). It may also be that quantum mechanics is not even a theory that gives us understanding (Cushing 1991), which would make attempts to accommodate quantum mechanics in an account of explanation questionable from the get go.

5 Explanatory demarcation: beyond the dichotomy

MLM-E has no truth requirement: the postulated entities (and their activities) are not required to be correct. An obvious worry is that the MLM-E account results in explanatory anarchism where "any explanation goes". While I reject explanatory anarchism, I am happy to embrace explanatory

[10] Note that the Duhem thesis seems also relevant to this point: theories are never applied to the world without a number of auxiliary assumptions about the experimental setup, the experimental apparatus, background effects, etc. Cartwright (1983) also speaks of a "prepared (i.e. idealized) description" of the phenomena which theories explain (rather than the nitty gritty of the actual world).

liberalism. In fact, I believe it's unavoidable: the truth requirement is not compatible with actual scientific practices regarding explanation.

Traditionally, accounts of scientific explanation used to require that the explanans be true for it to explain the explanandum phenomenon (Hempel 1965). As we mentioned earlier (Section 3.2), mechanistic accounts of explanation, too, require this. Yet the traditional dichotomy of "either truth, or no explanation (and understanding)" has been challenged in the past decade by whole trove of works (Elgin 2004, 2007, Strevens 2008, Batterman 2009, Bokulich 2011, Rohwer and Rice 2013, de Regt 2015, Potochnik 2017, Reutlinger et al. 2018, Doyle et al. 2019, Rice 2019). These works stress that explanations and understanding in science often are based on idealizations and literally false assumptions.

We thus face a dilemma: either we accommodate the explanatory practices of scientists and give up on the truth requirement, but then face the abys of explanatory anarchism, or we live on happily in denial of scientific practice. A possible solution of the dilemma lies in giving up on a *strict* truth requirement, without giving up on the truth requirement altogether.[11]

Approximate truth is a much used notion in the scientific realism debate. Attempts to formalize the notion have turned out difficult (see e.g. Chakravartty 2017), but the basic idea is clear enough: no realist believes that our best current scientific theories capture reality fully accurately, but only to some extent. Something along those lines might be said of scientific models: only those micro-level models are explanatory that capture the approximate truth about the postulated entities underlying micro-level explanations.[12]

Requiring only approximate truth rather than strict truth would help accommodate some of the examples discussed here: KT captures the approximate truth about gases consisting of molecules in motion (although it misrepresents many other features of molecules); Dalton's theory of atoms captures the approximate truth about chemical reactions consisting of the combination of atoms (although it misrepresents atoms as being indivisible); Mendelian genetics captures the approximate truth about inheritance being based on genes (although it oversimplifies inheritance).

In the realism debate, it is often disputed what parts of a theory were actually needed to generate empirically correct predictions (Vickers 2013). Likewise, a general worry about locating a model's explanatory power in its approximate truth might be that a model's idealizations are actually essential for its explanatory power. In KT, for example, taking into account the molecules'

[11] Some of the philosophers who made the above point have grappled with the issue in a different fashion. Strevens (2008) suggests that idealizations in reality just flag those causal factors that don't make a difference to the explanandum. Bokulich (2011, 2012) argues that model explanations are justified through "translation keys" with true theories. Pincock (forthcoming) defends the truth-requirement as necessary for explanation and further develops Strevens' view that idealizations have non-representational functions.

[12] Thanks to <blinded reference> for suggesting this to me. See also Pincock (forthcoming) for a 'veritist' proposal along those lines.

real properties, such as their extensions, does not result in the derivation of IGL, but rather in the derivation of the van der Waals equation (see also Section 2). The point has in fact been made quite extensively by several authors: idealizations are indispensable for certain explanations in physics and biology (Batterman 2009, Kennedy 2012, Rice 2015). For example, sometimes gases are falsely modeled as continuous fluids in order to account for 'shocks' in gases (i.e., areas of high molecular density within a gas created by external pressure). Without this idealization, Batterman (2009) contends, there wouldn't be any explanation of the target phenomenon. Similarly, Rice (2015) argues that false assumptions in "optimality explanations" in biology, such as the assumption that populations are infinite and that organisms mate randomly, cannot be removed from the model "without consequently eliminating the explanation being offered" (601).

If idealizations are indeed indispensable to model explanations, then the truth-requirement of explanations must go.[13] There is yet another reason for giving up on it: we would be in a much better position to make sense of why certain models in the past were adopted for their perceived ability to provide explanation and understanding.

Consider the notorious caloric theory of heat (CT), in which heat is conceived of as a *substance*, called caloric. Caloric particles are mutually repulsive, but attract (and are attracted by) matter particles (Figure 1).[14]

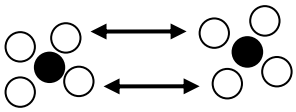

Figure 1: Caloric particles (white) repel each other but are attracted by matter particles (black).

Most interestingly for our purposes, CT can represent IGL and the associated counterfactuals as a necessity on the basis of its postulated entities (just like KT does; see Section 3). Consider for example this counterfactual supported by IGL: had the gas's volume been reduced, the pressure would have risen. According to CT, the counterfactual has this rather than another form (e.g. a pressure reduction upon the reduction of the gas's volume) because the mutual repulsion of caloric particles will be stronger the closer the particles are to each other (which will be the case when the volume is reduced). Or consider another counterfactual dependence: had the temperature of the gas increased (while holding fixed the volume of the gas), the pressure would have increased. On CT, the counterfactual has this rather than another form (e.g. a pressure increase upon a temperature increase), because an increase in the amount of caloric in the container (associated with an increase in temperature) necessarily leads to an increase in the net amount of repulsion between the caloric particles, which, in turn, is associated with an increase in pressure of gases on

[13] Some authors have argued that there are even models that employ fictions that presumably are not even partially true, such as the orbits in the Bohr model (Bokulich 2011). See Schindler (2014) and Nguyen (2020) for critical assessments.

[14] See e.g. Chang (2003) for more details.

CT. So it seems that CT, despite its patent falsity, can explain IGL in a very similar way as KT can: it too answers *ww*-questions.

One could of course simply take a conservative stance towards CT and insist that since caloric does not exist, CT did not explain at all. But, again, that would make the acceptance of CT by the contemporary scientific community at the time as an explanatory model a mystery (see e.g. Chang 2003). Alternatively, one could try to render the change from CT to KT one of structural continuity, despite a radical change in the postulated entities (Votsis and Schurz 2012). I am skeptical of such a rendering.[15]

In the spirit of explanatory liberalism, and against the dichotomy of "either truth or no explanation", I want to suggest that we simply accept that a model like CT is explanatory. At the same time there are very good grounds for thinking that KT is a *much better* explanatory model than CT. Those grounds have to do with the *theoretical virtues* of KT. For instance, KT has much wider *explanatory scope*, as it explains not only IGL, but also the properties of substances in different states, heat transfer and conduction of gases. Similarly, KT has made many successful predictions (such as the specific heat ratios of gases (de Regt 1996)) and has overall provided a very *fertile* research program (Clark 1976). In contrast, CT soon ran into problems that it could not solve, such as the apparently indefinite production of heat in the boring of cannons (as famously pointed out by Count Rumford) in contradiction with CT's central tenet that heat is a substance obeying the principles of conservation. Thus, by virtue of its wider explanatory scope and greater fertility, KT is a better explanation of IGL than CT.

Explanatory scope and fertility are of course only two out of a number of theoretical virtues on the basis of which scientists can assess their theories and models (Kuhn 1977, Schindler 2018). Simplicity and mathematical tractability are also important considerations, ruling out more fantastical 'models' with a much more demanding metaphysics than what's required by e.g. postulating particles and their interactions.

What emerges from these considerations is that scientific explanation is not a matter of "truth or no explanation", but rather more of a continuum on which there is a lot of differentiation between good and bad or no explanations.

6 Conclusion

Counterfactual dependence is widely seen as fundamental to scientific explanation. However in this paper I argued that there are kinds of explanation in science that take as their targets the regularities that support counterfactuals. These kinds of explanations therefore explain what counterfactual accounts of explanation take for granted, namely the particular form of the counterfactual. I call these explanations micro-level model explanations, or MLM-E, for short,

[15] The structure retained in this example seems a purely empirical structure we know by virtue of IGL. But structural realism requires structural continuity at the *theoretical* level. See Worrall (1989).

because they achieve their explanatory goals by postulating micro-entities and their activities. On the basis of these postulates, the model represents the contingent target regularities as necessities, thereby constrains the relevant counterfactuals, and provides answers to *why-would-things-have-been-different questions*.

References

Armstrong, David Malet. 1983. *What is a Law of Nature?* Cambridge: Cambridge University Press.

Batterman, Robert W. 2009. Idealization and modeling. *Synthese*, **169** (3): 427-446.

Bokulich, Alisa. 2011. How scientific models can explain. *Synthese*, **180** (1): 33-45.

———. 2012. Distinguishing Explanatory from Nonexplanatory Fictions. *Philosophy of Science*, **79** (5): 725-737.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Chakravartty, Anjan. 2017. Scientific Realism. *Standford Encyclopedia of Philosophy*, edited by Edward N. Zalta, <https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/>.

Chang, Hasok. 2003. Preservative realism and its discontents: Revisiting caloric. *Philosophy of science*, **70** (5): 902-912.

Clark, Peter. 1976. Atomism versus theromodynamics. In *Method and appraisal in the physical sciences: The critical background to modern science, 1800-1905*, Colin Howson (ed.), Cambridge: Cambridge University Press.

Colombo, Matteo, Stephan Hartmann, and Robert Van Iersel. 2015. Models, mechanisms, and coherence. *The British Journal for the Philosophy of Science*, **66** (1): 181-212.

Craver, Carl F. 2007. *Explaining the brain : mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Craver, Carl F. and James Tabery. 2019. Mechanisms in Science. *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, edited by Edward N. Zalta, <https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/>.

Cushing, James T. 1991. Quantum theory and explanatory discourse: endgame for understanding? *Philosophy of Science*, **58** (3): 337-358.

de Regt, Henk W. 1996. Philosophy and the Kinetic Theory of Gases. *The British Journal for the Philosophy of Science*, **47** (1): 31-62.

———. 2004. Review of James Woodward, making things happen. *Notre Dame Philosophical Review*. https://ndpr.nd.edu/reviews/making-things-happen-a-theory-of-causal-explanation/.

———. 2015. Scientific understanding: truth or dare? *Synthese*, **192** (12): 3781-3797.

de Regt, Henk W. and D Dieks. 2005. A contextual approach to scientific understanding. *Synthese*, **144** (1): 137-170.

Doyle, Y, S Egan, N Graham, et al. 2019. Non-factive understanding: A statement and a defense. *Journal of General Philosophy of Science*, **50** (3): 345-365.

Elgin, Catherine Z. 2004. True enough. *Philosophical issues*, **14** (1): 113-131.

———. 2007. Understanding and the facts. *Philosophical Studies*, **132** (1): 33-42.

Friedman, Michael. 1974. Explanation and scientific understanding. *The Journal of Philosophy*, **71** (1): 5-19.

Grüne-Yanoff, Till. 2013. Appraising models nonrepresentationally. *Philosophy of Science*, **80** (5): 850-861.

Hempel, Carl G. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Hempel, CG and P Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, **15** (2): 135-175.
Holton, Gerald James and Stephen G Brush. 2001. *Physics, the human adventure: From Copernicus to Einstein and beyond*. New Brunswick: Rutgers University Press.
Kennedy, Ashley Graham. 2012. A non representationalist view of model explanation. *Studies in History and Philosophy of Science Part A*, **43** (2): 326-332.
Kuhn, Thomas S. 1977. Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension*, Chicago: University of Chicago Press, 320-333.
Lange, Marc. 2017. *Because without cause: Non-causal explanations in science and mathematics*: Oxford University Press.
Nguyen, James. 2020. Do fictions explain? *Synthese*. https://doi.org/10.1007/s11229-020-02931-6.
Pincock, Christopher. forthcoming. A Defense of Truth as a Necessary Condition on Scientific Explanation. *Erkenntnis*.
Potochnik, Angela. 2017. *Idealization and the Aims of Science*. Chicago: University of Chicago Press.
Reutlinger, Alexander. 2016. Is there a monist theory of causal and noncausal explanations? The counterfactual theory of scientific explanation. *Philosophy of Science*, **83** (5): 733-745.
Reutlinger, Alexander, Dominik Hangleiter, and Stephan Hartmann. 2018. Understanding (With) Toy Models. *British Journal for the Philosophy of Science*, **69** (4): 1069–1099.
Reutlinger, Alexander and Juha Saatsi. 2018. *Explanation Beyond Causation: Philosophical Perspectives on Non-causal Explanations*: Oxford University Press.
Reutlinger, Alexander, Gerhard Schurz, and Andreas Hüttemann. 2017. Ceteris paribus laws. *Stanford encyclopedia of philosophy*, edited by Edward N. Zatla, <https://plato.stanford.edu/archives/spr2017/entries/ceteris-paribus/>.
Rice, Collin. 2015. Moving beyond causes: Optimality models and scientific explanation. *Noûs*, **49** (3): 589-615.
———. 2019. Models Don't Decompose That Way: A Holistic View of Idealized Models. *The British Journal for the Philosophy of Science*, **70** (1): 179-208.
Rohwer, Yasha and Collin Rice. 2013. Hypothetical pattern idealization and explanatory models. *Philosophy of Science*, **80** (3): 334-355.
Saatsi, Juha and Mark Pexton. 2013. Reassessing Woodward's account of explanation: Regularities, counterfactuals, and noncausal explanations. *Philosophy of Science*, **80** (5): 613-624.
Salmon, Wesley. 1984. *Scientific Explanation and Causal Structure of the World*. Princeton: Princeton University Press.
———. 1998. *Causality and explanation*. New York: Oxford University Press.
Schindler, Samuel. 2014. Explanatory fictions—for real? *Synthese*, **191** (8): 1741-1755.
———. 2018. *Theoretical Virtues in Science: Uncovering Reality Through Theory*. Cambridge: Cambridge University Press.
Strevens, Michael. 2008. *Depth: an account of scientific explanation*. Cambridge, Mass.: Harvard University Press.
Sullivan, Emily and Kareem Khalifa. 2019. Idealizations and Understanding: Much Ado About Nothing? *Australasian Journal of Philosophy*, **97** (4): 673-689.
Vickers, Peter. 2013. A Confrontation of Convergent Realism. *Philosophy of Science*, **80** (2): 189-211.

Votsis, Ioannis and Gerhard Schurz. 2012. A frame-theoretic analysis of two rival conceptions of heat. *Studies in History and Philosophy of Science Part A*, **43** (1): 105-114.

Woodward, James. 2003. *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.

———. 2018. Some Varieties of Non-Causal Explanation. In *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, Alexander Reutlinger and Juha Saatsi (eds.), Oxford: Oxford University Press.

Worrall, John. 1989. Structural Realism: The Best of Both Worlds? *Dialectica*, **43** (1-2): 99-124.