Erkenntnis (2021) 86:187–207
https://doi.org/10.1007/s10670-018-00101-2

ORIGINAL RESEARCH

# Metaphysical Explanation: The Kitcher Picture

Sam Baron[1] · James Norton[2]

Received: 29 May 2018 / Accepted: 24 December 2018 / Published online: 16 January 2019

**Abstract**
This paper offers a new account of metaphysical explanation. The account is modelled on Kitcher's unificationist approach to scientific explanation. We begin, in Sect. 2, by briefly introducing the notion of metaphysical explanation and outlining the target of analysis. After that, we introduce a unificationist account of metaphysical explanation (Sect. 3) before arguing that such an account is capable of capturing four core features of metaphysical explanations: (1) irreflexivity, (2) non-monotonicity, (3) asymmetry and (4) relevance. Since the unificationist theory of metaphysical explanation inherits irreflexivity and non-monotonicity directly from the unificationist theory of scientific explanation that underwrites it, we focus on demonstrating how the account can secure asymmetry and relevance (Sect. 4).

## 1 Introduction

This paper offers a new account of metaphysical explanation. The account is modelled on Kitcher's (1981/1989) unificationist approach to scientific explanation. We begin, in Sect. 2, by briefly introducing the notion of metaphysical explanation and outlining the target of analysis. After that, we introduce a unificationist account of metaphysical explanation (Sect. 3) before arguing that such an account is capable of capturing four core features of metaphysical explanations: (1) irreflexivity, (2) non-monotonicity, (3) asymmetry and (4) relevance. Since the unificationist theory of metaphysical explanation inherits irreflexivity and non-monotonicity directly from the unificationist theory of scientific explanation that underwrites it, we focus on demonstrating how the account can secure asymmetry and relevance (Sect. 4).

✉ Sam Baron
samuel.baron@uwa.edu.au

James Norton
james.norton@sydney.edu.au

[1] Department of Philosophy, School of Humanities, University of Western Australia, 35 Stirling Highway, Crawley 6009, Australia

[2] Department of Philosophy, University of Sydney, Sydney, Camperdown 2006, Australia

## 2 The Target Phenomenon

Consider the following claims:

(A) Sara exists *because* Sara's proper parts exist and are arranged a certain way.
(B) The apple is coloured *because* the apple is red.
(C) {Sara} exists *because* Sara exists.
(D) <Sara exists>[1] is true *because* Sara exists.
(E) The glass is fragile *because* the glass has crystalline bonds between its component molecules.
(F) Torture is wrong *because* torture causes a great amount of suffering.

Many metaphysicians have agreed with Rosen (2010) that claims like A–F exemplify a certain kind of non-causal explanation—often called *metaphysical explanation*—whereby in each case whatever follows the '*because*' is thought to explain (or partially explain) whatever precedes it. Metaphysical explanations are *non-diachronic*. Such explanations either involve states of affairs that obtain synchronically (like the apple being coloured and the apple being red) or they are cases where at least one object involved is an abstract object, and is thus not best thought of as existing at a time.

Most contemporary discussion of metaphysical explanation has been framed in terms of the recently popularised *grounding* relation.[2] There are, broadly, two ways of thinking about the relationship between metaphysical explanation and grounding. First, one might *identify* metaphysical explanation with grounding. According to such a view, the relation of grounding just *is* the relation of metaphysical explanation. Thus, to say that A metaphysically explains B is just to say that A grounds B. Second, one might hold that grounding is the relation that *backs* metaphysical explanation. According to the second view, grounding is to metaphysical explanation as causation is to scientific explanation. While a scientific explanation itself is just a set of propositions, what makes a set of propositions into an *explanation* is that it represents an underlying asymmetric determination relation, such as a causal relation.[3] Thus, on many contemporary theories of scientific explanation, causation is taken to be the relation in virtue of which a scientific explanation gets to be explanatory. Similarly, one might hold that what makes a set of propositions into a metaphysical explanation is that it represents an underlying grounding relation.[4]

[1] We use <P> to indicate the proposition that P.

[2] See, e.g., Schaffer (2009), Fine (2012), Audi (2012), Raven (2015) and Dasgupta (2014).

[3] See, e.g., Strevens' (2008) Kairetic theory of explanation.

[4] Such a view is particularly attractive if one believes that there is a strong analogy between grounding and causation. See Wilson (2017) and Schaffer (2016). Dasgupta (2017: 74) scathingly describes this as the view that grounding is "some metaphysical analogue of the Higgs boson that somehow [holds] the world together. The job of a metaphysician, on this […] conception, [is] to peer into reality and discern where these "groundons" [are] flowing (of course, to see these groundons one need[s] goggles provided by specialist departments)".

If one believes that the relation of grounding is to be identified with the relation of metaphysical explanation, then an analysis of metaphysical explanation amounts to an analysis of grounding.[5] If, however, one believes that the relation of grounding is *not* to be identified with the relation of metaphysical explanation, then an analysis of metaphysical explanation need not be an analysis of grounding. Of course, one might analyse metaphysical explanation in terms of grounding, just as an analysis of scientific explanation might appeal to causation. But metaphysical explanation might also be analysed without any appeal to a grounding relation whatsoever; just as one may seek to analyse scientific explanations without any appeal to causal relations (as the late positivists who were characteristically suspicious of metaphysical notions like 'causation' attempted to do).

Our goal in this paper is to provide an analysis of metaphysical explanation by extending a well-known theory of scientific explanation into the metaphysical domain. In particular, we aim to provide an account of why a given metaphysical explanation is explanatory *without* appealing to any grounding relations. Such a project is important for three broad reasons.

First, there are some who are sceptical of grounding.[6] Grounding, some maintain, simply does not exist; it is not a real relation. If grounding just is metaphysical explanation, then the view that grounding does not exist implies that there are no metaphysical explanations. Not every sceptic of grounding will be happy with that result. Some sceptics may doubt the existence of the grounding relation, but nonetheless believe that 'the apple is coloured *because* the apple is red' is a good explanation. By offering a viable analysis of metaphysical explanation that makes no use of grounding relations, we make room for this kind of view.[7]

Second, even if one is not sceptical of grounding relations, one may still think that grounding and metaphysical explanation are different phenomena, that are deserving of different treatment. By developing such an account, we widen the space of theoretical options, and open up the possibility of separating grounding from metaphysical explanation.

Third, by extending a theory of scientific explanation into the metaphysical domain, we can potentially unify the two kinds of explanation together. This is an attractive prospect. It would be very pleasing to have a broad theory of explanation that applies to explanations no matter where they arise. Such a theory would boast the benefits of elegance and power. The first step in developing such a theory is to take each existing account of scientific explanation and attempt to generalise it.

A theory of metaphysical explanation will tell us what it is for one fact to metaphysically explain another.[8] Such a theory is *adequate* just to the extent to which it fits the target phenomenon by identifying all and only those sets of propositions that

[5] For one recent attempt to provide a theory of metaphysical explanation along these lines, see Wilsch (2015/2016).

[6] See Daly (2012) and Wilson (2014) for critiques of grounding. See Raven (2012), Audi (2012) and Rodriguez-Pereyra (2005) for defences of grounding.

[7] We are not the first to seek a theory of metaphysical explanation that does not make use of grounding relations. See Shaheen (2017), Norton and Miller (2017) and Thompson (2018).

[8] Note that by 'fact' we mean 'true proposition' rather than 'state of affairs'.

we pre-theoretically take to be metaphysical explanations as in fact being metaphysical explanations. It is a large task to show that every putative metaphysical explanation can be captured by a given theory, and so it is doubtful that we can demonstrate the adequacy of our preferred theory in this paper.

Instead, we will focus on achieving two more modest goals. Our first goal is to explicitly formulate an alternative to grounding-based theories of metaphysical explanation. The option we propose is based on Kitcher's unificationist theory of scientific explanation. We will reiterate how Kitcher's (1981/1989) theory works in the scientific context before adapting Kitcher's account to provide a theory of metaphysical explanation. Our second goal is to show that the application of the unificationist machinery yields explanations that are irreflexive, non-monotonic, asymmetric and constrained by relevance—four core features of metaphysical explanations—thereby demonstrating that grounding-based accounts are not the only ones to do so. Since the unificationist theory of metaphysical explanation inherits irreflexivity and non-monotonicity directly from the unificationist theory of scientific explanation, we focus on demonstrating how the account can secure asymmetry and relevance. Along the way, we will apply the unificationist machinery to several intuitive cases of metaphysical explanation, thereby demonstrating how to capture metaphysical explanations within the framework. By showing that the account captures core features of the notion of metaphysical explanation, we aim to show that it is a viable contender, and thus worthy of being explored in greater detail in the future.

## 3 Unification

Let us turn now to the project of adapting Kitcher's theory of explanation for use in a metaphysical context. We will start, as noted, with an overview of Kitcher's theory of scientific explanation.

### 3.1 Scientific Unification

Kitcher's theory is best understood as an extension of the classic deductive-nomological (DN) model, one that emphasises the role of unification in making derivations explanatory. Kitcher constrains the class of DN derivations that count as explanatory by allowing only those that best unify that which needs to be explained. Thus, one must consider one's entire corpus of beliefs holistically before one can know what explains what. Those derivations that feature in the best systematisation of the DN derivations are the explanatory derivations. In brief, the best system is that which uses the fewest argument patterns to generate the biggest conclusion set, while keeping the patterns stringent, such that derivations which instantiate the same argument pattern are genuinely similar. The resulting theory has the power to rule out some of the DN derivations which do not strike us as explanatory.

For Kitcher, a derivation is an ordered pair: a set of statements to serve as premises, and a conclusion statement. We take the set of statements/propositions from which these ordered pairs are built to be consistent and deductively closed, and name the set K. K is a systematisation of (a consistent version of) the set of statements endorsed by the scientific community. There is a set of derivations which best unifies K. This is known as the explanatory store over K, or E(K). To be an explanatory derivation is simply to be a member of E(K). The subset of sentences in E(K) that are conclusions of the derivations therein is known as C(E(K)).

Given a certain K, there will be many candidate E(K)s just as there are many ways to systematise the same corpus of beliefs. Crucial to the unification theory is that not all of these systematisations are equal. There is a privileged set of derivations which best unifies K, and it is only these derivations which qualify as explanations. In order to develop criteria that allow us to compare two attempts to unify K [that is, two candidate E(K)s] we require the notion of a general argument pattern, such that it is clear when two explanations share a pattern, and clear how many different patterns are being used to systematise a corpus of beliefs.

General argument patterns are a kind of schematic argument, built from schematic sentences. Following Kitcher's (1989) example, consider the sentence:

Organisms homozygous for the sickling allele develop sickle-cell anaemia.

We can turn this sentence into a schematic sentence by replacing some variables with dummy letters as follows:

Organisms homozygous for A develop P.

Unlike the original, this schematic sentence (appropriately filled) tells us about a variety of relationships between homozygous genotypes and particular phenotypes. Yet the variables are restricted in the kinds of things which can acceptably fill them. A is to be filled with an allele, and P with the corresponding phenotype. No other substitutions are permitted. Thus, schematic sentences must be paired with instructions on how to restrict substitutions for the variables. These restrictions on substitutions are called *filling instructions*, and they ensure that the unificationist's patterns are appropriately constrained.

Note that a sentence can be schematised to greater and lesser degrees. The highest degree of schematisation involves replacing each non-logical expression in a sentence with a dummy letter. The filling instructions can then be modified to specify the substitutions that are allowable for each dummy letter. Note, however, that the filling instructions for any dummy letter can be specified so tightly that only *one* thing can be allowably substituted for the dummy letter. This means that we can take a maximally schematised sentence and, by tightly specifying the filling instructions for one or more of the schematised expressions, produce a sentence that is, for all intents and purposes, equivalent to a less schematised version of the relevant sentence. So, for example, suppose we take the sentence 'organisms homozygous for the sickling allele develop sickle-cell anaemia' and schematise all of the non-logical expressions to produce something like:

O's H for A D P.

We can then specify in the filling instructions for this sentence that O can *only* be an organism, H can *only* be homozygosity and D can *only* be development. We can, however, allow that the filling instructions for A and P correspond to the filling instructions for the semi-schematized version of the sentence considered above. This means that the two schematized sentences are equivalent, in so far as the instances of the sentences are concerned: they allow all and only the same instances. As we shall see in a moment, one of the primary dimensions of variation for argument patterns concerns how tightly constrained the filling instructions are for the sentences in that pattern. Altering the degree of schematisation is, in some sense, a special case of altering the constraints that are placed on a sentence by the filling instructions (and so, as we shall see, degree of schematisation is a special case of *stringency*).

A schematic argument, then, is a set of schematic sentence/filling instruction pairs like the above. A schematic argument is accompanied by a *classification* which describes the inferential characteristics of the argument. That is, the classification tells us which sentences are premises, which conclusions, and how we can infer some from others.

A general argument pattern is a schematic argument (complete with filling instructions for each schematic sentence) and a classification. Kitcher tells us that a particular derivation instantiates a general argument pattern if:

1. The sequence has the same number of terms as the schematic argument of the general argument pattern.
2. Each sentence in the sequence is obtained from the corresponding schematic sentence in accordance with the appropriate set of filling instructions.
3. It is possible to construct a chain of reasoning which assigns to each sentence the status accorded to the corresponding schematic sentence by the classification. (1981: 517)

Thinking in terms of general argument patterns provides a useful framework within which to think about similarities between derivations. Derivations are similar to one another in virtue of instantiating a common general argument pattern. Any derivation is maximally similar to itself, and likely has some similarity of form to any derivation (perhaps they are all deductively valid, for example). Yet such similarity admits of degrees. A pair of derivations is more similar than another pair if the general argument pattern they both instantiate is more stringent. Stringency is determined by the extent to which we restrict the filling instructions and classification. Maximal stringency leads to a single-case 'pattern' (a general argument pattern so tightly constrained as to only cover one particular derivation), whilst minimal stringency allows the other degenerate case of the all-inclusive pattern (a general argument pattern so loosely constrained as to cover all derivations).

To summarise thus far, we have multiple competing sets of derivations [explanatory stores; candidate E(K)s], each of which is viable for the derivation of members of K. The members of each candidate E(K) instantiate some general argument patterns. Thus, we can say that each candidate E(K) is ‘backed’ by a store of general argument patterns. There will be variety in the number of patterns a certain candidate E(K) makes use of, and in the stringency of those patterns. The ‘backing’ of a certain candidate E(K) is its *generating set*. Various candidate E(K)s, each aiming to be the explanatory systematisation of K, can be compared on the basis of their respective generating sets.

For example, one candidate E(K) might make use of 20 general argument patterns to derive its conclusion set, while another might only use 15 general argument patterns to derive the same conclusions. The former uses more patterns, but the patterns are, let us suppose, more stringent. The latter uses fewer patterns, but the patterns are less stringent. Kitcher provides two competing criteria, such that the best systematisation of K will be the candidate E(K) which strikes the best balance between the two.

The first criterion is that of paucity of patterns. We will sometimes refer to this simply as ‘paucity’. This criterion tells us that more unification is achieved through deriving as many conclusions as possible from the fewest number of argument patterns. Thus, a small generating set with a large conclusion set is better, all else being equal. The goal is to use few, powerful patterns to generate as many conclusions as possible. A motivation for this criterion is that a systematisation that uses two different argument patterns in similar cases where a single pattern could have done the job has failed to unify.

Paucity ensures that explanations are non-monotonic (i.e. ensures that arbitrary premises cannot be added to an explanatory derivation and the resulting derivation still count as an explanation). To see why, take any argument pattern P that appears inside a candidate E(K). Another argument pattern can be produced by taking P and adding a further premise A into the premise set. An instance of any such pattern will be valid if the corresponding instance of P is valid. Because we can add any proposition to P whatsoever and produce another valid pattern, this gives us a series of patterns to consider, each of which involves adding a slightly different proposition to P. Call these the (P + A) patterns.

If P and all of the (P + A) patterns are part of the candidate E(K) that best systematises our beliefs, then we will be able to take any instance of P that is valid, and add any proposition to that argument whatsoever and still have a derivation which is part of E(K), and is thus an acceptable explanation. This, in turn, will force the explanation at issue to be monotonic. To ensure the non-monotonicity of the explanation, then, it must be shown that it is not the case that both P and all of the (P + A) patterns are a part of the generating set. This is ensured by paucity. Given paucity, only one of these patterns—be it P or one of the (P + A) patterns—will be in the generating set. That is because, as noted, if a candidate E(K) makes use of two patterns to derive the same conclusions as those which another candidate E(K) can derive using one pattern, the former candidate E(K) has failed to unify. The P pattern and each of the (P + A) patterns enable us to derive the same things, and so only one of these patterns will be a part of the generating set for E(K).

The second criterion is that of stringency of patterns. We will sometimes refer to this simply as ‘stringency’. This criterion tells us that if we go overboard with paucity of patterns such that the general argument patterns in the generating set are insufficiently stringent, the unification will not be genuine. Recall that stringency is determined by the strictness of the filling instructions and the classification. Thus, the second criterion serves to constrain the first. If all that mattered was minimising the number of argument patterns, this would grant favour to candidate E(K)s which use a single, unilluminating pattern for every derivation, and thus achieve merely ‘spurious’ unification.

Stringency also ensures that there are no reflexive derivations in E(K): no cases in which a single proposition is used to derive itself. As noted, stringency is partly a measure of the degree of unification between the conclusions that may be derived from a given general argument pattern. The paradigm case of a failure of stringency whereby merely spurious unification is achieved is the case of Kitcher’s (1981) God pattern: God wills that F, therefore F. A reflexive argument pattern is no better than the God pattern. If any reflexive pattern of the form: P, therefore P (where any proposition can be substituted in for P) is allowed into the generating set then it will be possible to use that single pattern to derive anything, which clearly won’t do.[9] Thus the unificationist machinery delivers the result that there are no reflexive explanations.

The difference between ‘spurious’ and ‘genuine’ unification is difficult to state exactly. One way to specify the difference is via similarity. A case of genuine unification is one in which the conclusions that are unified by a given argument pattern appear to be relevantly similar to one another: there is some feature held in common between the conclusions and it is in virtue of this similarity that the conclusions fall under the argument pattern that they do. This similarity of the conclusions in turn ought to mirror a similarity in the pattern itself. What we want, ideally, is for there to be some feature held in common between the conclusions in virtue of which those conclusions can all be derived from similar premises in a similar way. A case of spurious unification is one in which one or more of these dimensions of similarity is absent.

Kitcher provides the following corollaries:

(A) Let $\Sigma$, $\Sigma'$ be sets of arguments acceptable relative to K (i.e. potential E(K)s) which meet the following conditions:

  (i) the basis of $\Sigma'$ is as good as the basis of $\Sigma$ in terms of the criteria of stringency of patterns, paucity of patterns, presence of core patterns, etc.

  (ii) $C(\Sigma)$ is a proper subset of $C(\Sigma')$ [recall that $C(\Sigma)$ is the set of the conclusions of arguments in $\Sigma$].

  Then $\Sigma \neq E(K)$.

[9] It might be objected that reflexive argument patterns can be rendered more stringent via restrictions on their filling instructions. For instance, one could restrict the pattern: P therefore P to only range over a single proposition, or a class of propositions. This move is pre-empted by Kitcher, who objects that in such patterns all the work is being done by the filling instructions, and the nonlogical vocabulary in the premises is idle. Thus, any apparent unification offered by patterns of self-derivation is spurious because the non-logical vocabulary ought to be contributing to the unification provided by the pattern (1981: 526–529).

(B) Let $\Sigma$, $\Sigma'$ be sets of arguments acceptable relative to K [i.e. potential E(K)s] which meet the following conditions:

> (i) $C(\Sigma) = C(\Sigma')$.
> (ii) The basis of $\Sigma'$ is a proper subset of the basis of $\Sigma$.
>
> Then $\Sigma \neq E(K)$.
> (1981: 522; slightly altered for readability).

Corollary B formalises the criterion of paucity of patterns: if two E(K)s have equal sized conclusion sets we should prefer the one with a smaller generating set. Corollary A formalises the further criterion that all else being equal, an E(K) with a larger conclusion set is preferable, as it explains more phenomena.

Paucity and stringency give rise to a ‘Goldilocks’ trade-off, which concerns precisely how to strike the right balance between them. We want to derive as much as possible from the fewest number of patterns, yet those patterns must be stringent enough that instances of the same pattern are genuinely similar. Kitcher doesn’t provide a general rule which governs the balancing of these conditions: we must decide on a case-by-case basis.

In addition to corollaries A and B, Kitcher adds a further *minimality* condition. To see why there is a need for such a condition, consider again the P pattern and the (P + A) patterns considered above. As noted, paucity ensures that only one of these patterns will be a part of the generating set for E(K). But which one? Intuitively, it is the P pattern that we should include, since each of the (P + A) patterns contains redundant information. But paucity and stringency alone cannot deliver this result. For suppose that we have three candidate E(K)s: $\Sigma$, $\Sigma^*$ and $\Sigma'$. The only difference between these three candidate E(K)s is that $\Sigma'$ contains both of the argument patterns just described, $\Sigma^*$ contains just (P + A) and $\Sigma$ contains just P.

First, Kitcher’s Corollary B tells us that $\Sigma$ and $\Sigma^*$ are better than $\Sigma'$: if two candidate E(K)s have equal sized conclusion sets we should prefer the one with a smaller generating set. To see this, note that $\Sigma$, $\Sigma^*$ and $\Sigma'$ all have the same sized conclusion sets (indeed, the conclusion sets are exactly the same). $\Sigma$ and $\Sigma^*$, however, have smaller generating sets than does $\Sigma'$. $\Sigma$ and $\Sigma^*$ have equally sized generating sets, however, and thus Corollary B is indifferent between them. Thus, in order to select between a candidate E(K) that makes use of pattern P and one that makes use of pattern (P + A), we require a minimality condition. This condition says that for any two patterns that are otherwise matched with respect to stringency, and that have the same conclusion sets, we should choose the pattern with the smaller premise set (Kitcher 1981: 524).

In sum, the unificationist view of scientific explanation differs from the DN view in that the status of a particular derivation as explanatory (or not) cannot be assessed in isolation. Rather, it must be evaluated as part of a system of derivations. These systems can be compared on the basis of the number of general argument patterns they instantiate, the stringency of these patterns, and the conclusions generated. The best system will strike the right balance for this Goldilocks trade-off, such that few patterns are used, yet the patterns are such that their instantiations are genuinely similar. The derivations that are part of this best system are the explanations.

## 3.2 Metaphysical Unification

Let us now transpose Kitcher's account into a metaphysical key. The machinery remains the same. The only difference is with K. On Kitcher's view, K is constituted by those beliefs that are important for scientific explanations only. Beliefs about what observable phenomena there are, for example, along with beliefs about what the laws of nature are. In the metaphysical case, the aim is to unify $K_M$, which is a different set of beliefs. The set $K_M$ is constituted by those beliefs that are implicated in metaphysical explanations, such as beliefs about parts and wholes, the existence of sets and their members, and dispositional and categorical properties. K and $K_M$ are not disjoint sets. The two sets are partially overlapping. That's because at least some metaphysical explanations involve beliefs about the kind of phenomena that appear in scientific explanations. Consider, for instance, the existence of some molecule (call it molecule M). M's existence will plausibly be *metaphysically* explained by the existence of M's constituent atoms, but *scientifically* explained by the physical processes that combined those atoms at an earlier time. Thus, the sentence 'Molecule M exists' will appear in both K and $K_M$.

For our purposes, it will be enough if $K_M$ contains all of the statements that are supposed to be involved in metaphysical explanations, wherever they arise. What we are looking for, then, is $E(K_M)$. Those derivations that constitute $E(K_M)$ are the metaphysical explanations. A version of the unificationist account of metaphysical explanation can therefore be stated as follows:

> **The Unificationist Account**
> $\Gamma \vdash \Delta$ is a metaphysical explanation iff there is an argument pattern in the generating set for $E(K_M)$ such that $\Gamma \vdash \Delta$ is an instance of that pattern.[10]

The unificationist account analyses the relation in virtue of which one fact metaphysically explains another as a relation of derivation. It is not just any relation of derivation that will do, however. Rather, metaphysical explanations are those derivations that instantiate an argument pattern that is a member of the generating set that best unifies our metaphysical beliefs.

It is worth noting that our unificationist account is simplified in the following respect. We are seeking an $E(K_M)$ that unifies statements that appear in metaphysical explanations only (some of which appear in scientific explanations). The alternative would be to start with the beliefs that are implicated in all scientific and metaphysical explanations and then unify across the entire set of beliefs. Indeed, this broader kind of unification would seem to fit better with the spirit of Kitcher's unificationist outlook. As we have seen, Kitcher's unificationism aims to look at explanation in a holistic manner. Explanations are never assessed in isolation; they fit within a

[10] Γ and Δ are sets of sentences.

pattern of explanatory inferences. This holistic approach speaks in favour of unifying all explanations together, wherever they ultimately arise, which is something that the unificationist account we have outlined for metaphysical explanation does not yet do.

We are prepared to accept as a limitation of the current proposal that it cleaves scientific from metaphysical explanation. Our goal is to get a basic version of the theory up and running. Unifying across science and metaphysics will introduce unwanted complexities, at least at this stage of the theory-building process. We also want to leave it open that Kitcher's account is apt for metaphysical explanation but that it may not work for scientific explanation. Unifying scientific and metaphysical explanation together from the outset would seem to foreclose that possibility a bit too early. That being said, we aim to develop the view in such a manner that it can be appropriately unified across science and metaphysics, which is what we ultimately want the theory to do.

It is more or less trivial to show that each of the cases of metaphysical explanation outlined in Sect. 2 can be arranged into deductive argument form. Thus formulated, it is then straightforward to schematise the arguments. So, we will largely leave this as an exercise for the reader. In what follows, however, we will need some patterns to work with, and so we will formalise some cases as we go.

## 4 Asymmetry and Relevance

The unificationist account of metaphysical explanation rules out irreflexive and monotonic explanations in precisely the same way as does the unificationist account of scientific explanation (via stringency and paucity respectively). Thus, in order to demonstrate the viability of the theory our goal is to show that, when applied to metaphysical explanations, the unificationist machinery has the capacity to yield the asymmetry and relevance of those explanations. In quite general terms, an explanation is *asymmetric* when if A explains B it is not the case that B explains A. An explanation is *relevant*, by contrast, when if A explains B only information that is relevant to B appears in A.

Let us begin with asymmetry. Consider a particular singleton set {2} and its urelement 2. It seems correct to say that if {2} exists then 2 exists. Equally, if 2 exists then {2} exists. The two entities mutually necessitate each other. There is therefore a deductive symmetry between the two facts. The symmetry can be captured via two general argument patterns. First, we can formulate a *singleton set formation pattern*:

(1) E exists.
(2) Necessarily, for any entity E, E exists just in case the singleton set {E} exists.

Therefore,

(3) {E} exists.

The filling instructions tell us that any entity can be substituted for E. The classification tells us that (3) follows from (1) and (2) by modus ponens. Here is the *urelement formation pattern*:

(1) {E} exists.
(2) Necessarily, for any entity E, E exists just in case the singleton set {E} exists.

Therefore,

(3) E exists.

The filling instructions tell us that any entity can be substituted for E. The classification tells us that (3) follows from (1) and (2) by modus ponens.

Clearly, we want to allow that only one of these argument patterns is a part of the generating set for E($K_M$). If we allow both, then there will be cases of symmetrical derivation of sets from urelements and back again, and thus there will be cases of symmetrical explanation in E($K_M$), which will violate a plausible constraint on metaphysical explanation, namely that it is asymmetric.

In a moment, we will show how the unificationist theory can exclude symmetrical explanations. First, though, we will outline a case of irrelevance since the same solution as the one used for symmetry can be used there as well. Consider the following two general argument patterns. First, the *spurious number pattern*:

(1) E exists.
(2) If E exists then N exists.

Therefore,

(3) N exists.

The filling instructions tell us that any entity can be substituted for E, and any number can be substituted for N. The classification tells us that (3) follows from (1) and (2) by modus ponens.

Second, the *spurious universal pattern.*

(1) E exists.
(2) If E exists then U exists.

Therefore,

(3) U exists.

The filling instructions tell us that any entity can be substituted for E, and any universal can be substituted for U. The classification tells us that (3) follows from (1) and (2) by modus ponens.

Clearly, we wouldn't want either of these spurious patterns to be part of the generating set for E($K_M$). The first allows us to derive the existence of any number from the existence of anything whatsoever. So, for instance, we can derive the existence of the number 2 from the fact that Sara has a big ginger cat. But surely Sara's big ginger cat does not explain why the number 2 exists. Sara's big ginger cat is completely irrelevant to the number 2. Similarly, we can derive the existence of the universal *blueness* from the existence of the same big ginger cat, which is absurd. In short, these spurious patterns allow too many, irrelevant derivations to count as metaphysical explanations.

There are two broad tools that the unificationist has at her disposal to address the threat of symmetry and irrelevance. The first is *stringency*. Consider the set formation and urelement formation patterns. Sets are cheap: for any object, there is a singleton set containing that object. Accordingly, the urelement formation pattern can be used to derive the existence of any object whatsoever. This, like the reflexive pattern discussed in Sect. 3.1, is a case of spurious unification. The conclusions that are being unified under the urelement formation pattern have little in common with one another. And even if there are commonalities among some of the elements (for instance, a large group of elements are numbers) those commonalities do not explain why it is that the conclusions in question can be derived using this pattern.

Compare this to the set formation pattern. For one thing, the set formation pattern can only be used to derive the existence of sets, and is thus more stringent than the urelement formation pattern. The unification achieved by using this particular pattern does not appear to be spurious. The conclusions that are unified all have something in common—they are all about sets—and it is clear why conclusions of this type can be derived from this particular argument pattern. The pattern is a pattern regarding set formation, and that's why facts about sets are unified by the pattern in question.

Stringency can also be used to handle the relevance problem. Consider the spurious universal and spurious number patterns. At first glance, stringency might not seem to help. The conclusions in both cases are appropriately unified. In the spurious number pattern, all of the conclusions are about numbers, and so appear to be unified under this commonality. In the spurious universal pattern, all of the conclusions are about universals, and so appear to be unified under this commonality.

However, as discussed, stringency is not just about unifying the conclusions. This unification must be connected back to the derivation itself. The conclusions need to be appropriately unified, where it is evident that this unification is explained by the nature of the general argument pattern. But these general arguments patterns don't explain the unification in question. Far from it: it is quite mysterious from the argument pattern as to why numbers or universals are being unified in this way. A better general argument pattern would link the unification of the conclusion to some feature that numbers or universals all have in common; a feature that is used to drive the derivation.

In other words, while it is true that the conclusions of the spurious universal and number patterns all have something in common, the patterns themselves are not sufficiently stringent. We can see this by considering the fact that there is no obvious similarity across the premises of the derivations that are instances of the broad patterns. Both patterns use the troubling premise: (1) E exists, where any entity can be substituted for E. As a result, the derivations that are instances of the pattern will be quite heterogeneous with respect to the kinds of entities that are mentioned in the relevant patterns. Accordingly, when we look across the derivations that are instances of these spurious patterns, there is no way to link the commonality in the conclusions to some commonality in the premises. In both cases, the space of conclusions may be unified, but the space of premises is not: it is a hodgepodge.

What we are looking for, then, is a general argument pattern with a conclusion set that is unified in virtue of some commonality between the various ways the pattern can be filled. For an example of such a common feature, consider that it might be that there is a certain property that all fragile objects share—such as the disposition to break—and the possession of this property by those objects is explained by the fact that those objects each possess one of a relatively small group of microphysical structures. An argument pattern that connected dispositional properties—such as the disposition to break—to the categorical properties upon which they are based would better satisfy the stringency constraint. It wouldn't merely unify the conclusions by being able to derive them all, it would unify them all *in virtue of* some important feature that they have in common.

This leads us nicely to a second way to handle the symmetry and relevance problems, by appealing to the *paucity of patterns*. To see how this constraint works in the symmetry case, it is useful to consider a slightly different example. Consider the following two argument patterns. First, the *categorical–dispositional pattern*:

(1) E has a categorical property of kind K.
(2) Necessarily, for any entity E, E possesses a categorical property of kind K, just in case E possesses dispositional property D.

Therefore,

(3) E possesses dispositional property D.

The filling instructions tell us that any entity can be substituted for E, and that the K-properties are those categorical properties that realise disposition D. The classification tells us that (3) follows from (1) and (2) by modus ponens. So, for instance, appropriately filled, this pattern can derive that an object is fragile from the fact that it possesses the appropriate microphysical structure.

Next, the *dispositional–categorical pattern*:

(1) E possesses dispositional property D.
(2) Necessarily, for any entity E, E possesses a categorical property of kind K, just in case E possesses dispositional property D.

Therefore,

(3) E has a categorical property of kind K.

The filling instructions tell us that any entity can be substituted for E, and that the K-properties are those categorical properties that realise disposition D. The classification tells us that (3) follows from (1) and (2) by modus ponens.

The categorical–dispositional pattern looks like the kind of pattern that we might want to be a part of the generating set. But will the dispositional–categorical pattern be part of the generating set? Well, according to paucity, we should only make use of this pattern if it can generate new conclusions. Given that the conclusions of this pattern tell us that some entity has a certain determinable property (like having a property of kind K), it is likely that the conclusion set will be a subset of the conclusion set of another pattern that we have good reason to suppose is part of the generating set, namely: the *determinate–determinable pattern*. This pattern can be set out as follows:

(1) For any entity E, E has determinate property P only if E has determinable property Q.
(2) E has determinate property P.

Therefore,

(3) E has determinable property Q.

The filling instructions tell us that any entity can be substituted for E, P is a determinate property and Q is a determinable of P. The classification tells us that (3) follows from (1) and (2) by modus ponens.

This pattern derives that objects have determinable properties on the basis of their determinate properties. So, for example, we can capture the claim that the apple is coloured *because* the apple is red by substituting the apple for E, redness for P and colouration for Q. Importantly, the conclusion set of the dispositional–categorical pattern is a mere subset of the conclusion set of the determinate–determinable pattern because the dispositional–categorical pattern will only derive that objects have those determinable properties that are the basis of some disposition. Thus, paucity of patterns tells us to jettison the dispositional–categorical pattern in favour of the more powerful determinate–determinable pattern.

Similar broad considerations apply to the spurious number and spurious universal patterns. In both cases there are other patterns that are capable of generating the same conclusions, but that are ultimately more powerful. In the case of numbers, numbers are not the only mathematical entities whose existence we may want to explain. The existence of sets, functions, classes, groups and so on may all stand in need of explanation. Some of these may be inexplicable. That's fine. But many will need to be explained, and so will need to be unified within E($K_M$). It is plausible to suppose that there is a general argument pattern that allows us to derive, say, the existence of all numbers and functions from the existence of sets. On some accounts of the foundations of mathematics, sets are responsible for both numbers and functions, and so we can imagine there being some feature held in common between these two mathematical kinds that, on the one hand, is based in set theory and, on the other hand, is a feature in virtue of which they may be unified under a stringent common pattern.

A good candidate is the notion of *structure*: numbers and functions both have a particular kind of mathematical structure which, according to some mathematicians, can be revealed set-theoretically. If that's right, however, then paucity will demand that we include this more general argument pattern—the pattern that takes in sets and yields both numbers and functions—rather than a pattern that yields numbers only, as does the spurious number pattern. It is less clear how this second solution applies to the spurious universal pattern. That's because it is less clear in general what metaphysically explains universals full stop, and so it is unclear what else that thing might explain, and thus what a more general argument pattern might look like in this case.

So far, we have argued that symmetry problems can be dealt with using stringency, paucity or, indeed, some combination of the two. There is, however, a version of the symmetry problem that warrants further attention.[11] For it is not obvious exactly how stringency and paucity alone can deal with this further incarnation of the worry.

To see the problem, consider the following *composition pattern*.

(1) Relation R obtains between the collection of entities $P_1 \ldots P_n$.

(2) Necessarily, for any entities $P_1 \ldots P_n$, if the $P_n$ exist and relation R obtains between the $P_n$, then W exists.

Therefore,

(3) W exists.

The filling instructions tell us that the $P_n$ are non-overlapping concrete objects, R tells us how the $P_n$ must be arranged in order for there to exist a composite whole, W, composed by the $P_n$. The classification tells us that (3) follows from (1) and (2) by modus ponens.

This pattern allows us to derive the existence of a composite object via the arrangement relation that obtains between its parts. So, for example, we can capture the claim that Sara exists because Sara's proper parts exist and are arranged a certain way by substituting Sara in for W in the schema; Sara's parts in for the $P_n$ and the arrangement relation in for R.

But now consider the following *modus tollens composition pattern*:

(1) It is not the case that W exists.

(2) Necessarily, for any entities $P_1 \ldots P_n$ if the $P_n$ exist and relation R obtains between the $P_n$, then W exists.

[11] This version of the problem is discussed by Wilsch (2016) in the context of developing a DN theory of metaphysical explanation.

Therefore,

(3) It is not the case that relation R obtains between the collection of entities $P_1$… $P_n$.

The filling instructions tell us that the $P_n$ are non-overlapping concrete objects, R tells us how the $P_n$ must be arranged in order for there to exist a composite whole, W, potentially composed by the $P_n$. The classification tells us that (3) follows from (1) and (2) by modus tollens.

Instances of the first pattern seem to count as explanations. If these count as explanations, however, then one might not want instances of the second pattern to count as explanations. If the first pattern is in the generating set, then this seems to suggest that the direction of explanation moves from arrangement relations between parts to the existence of wholes. The second pattern, however, explains the (non-existence) of a certain arrangement relation in virtue of the (non-existence) of the whole. This, in turn, suggests that the direction of explanation moves from facts about wholes to facts about parts being arranged in a certain way. If we are forced to accept both patterns in our generating set then we are forced to accept a certain kind of coarse-grained explanatory symmetry: wholes explain facts about parts and vice versa. In so far as this kind of symmetry is objectionable we require a further reason for thinking that only the modus ponens composition pattern is part of the generating set for $E(K_M)$.

Note that it is by no means guaranteed that the modus tollens composition pattern will be a part of the generating set when the modus ponens pattern is a part of the generating set. That's because it may turn out that, on the one hand, there is no general argument pattern that better unifies $K_M$ than the modus ponens pattern; while there *is* a general argument pattern that better unifies $K_M$ than the modus tollens pattern. How might this happen? Well, it may turn out that the modus ponens pattern is the best way to derive facts about the existence of wholes within the competing constraints of stringency and paucity, whereas the modus tollens pattern is *not* the best way to derive facts about the non-existence of metaphysically relevant relations; there is another pattern that better satisfies the trade-off between the stringency and paucity of patterns.

The question, then, becomes: how else might we derive facts about the non-existence of metaphysically relevant relations between parts or the non-existence of the parts themselves? Paucity and stringency can be used to force this kind of choice, but to fully solve the problem we need to also provide a better pattern; one that outcompetes the modus tollens pattern. So, paucity and stringency alone won't help to solve every putative case of symmetry.

Before we consider the question of how we might best the modus tollens pattern, it is useful to consider a prior question: do we even *want* to explain the non-existence of relations between parts? One might hold the view that facts about the *non-existence* of certain metaphysically relevant relations don't require explanation at all. These facts may simply not be inside $K_M$. One might hold this view if one thinks that metaphysical explanation is restricted to explaining the presence of something in terms of the presence of something else; metaphysical explanation is not in the business of explaining absences.

Such a view would be very convenient for our purposes. The modus tollens composition pattern would not be a part of the generating set for E($K_M$) because we are not trying to unify beliefs about the non-existence of metaphysically relevant relations. But while we can see that some may be attracted to such a view, we deem it to be unduly restrictive. We see no reason why metaphysical explanation cannot be a matter of explaining an absence, and thus we take seriously the idea that the conclusion of the modus tollens composition pattern is the kind of thing we may want to explain.

This returns us to the question of how else we might derive facts about the non-existence of arrangement relations between parts. In the end, this could be done in a number of ways, depending on what, exactly, the relation R in the composition pattern and its modus tollens cousin is supposed to be. But to get a feel for the kind of solution we endorse, it is useful to consider a particular example. Suppose that R is a relation of *spatiotemporal contiguity*. Then the explanandum in this case is the fact that the $P_n$ do not stand in a relation of spatiotemporal contiguity with one another. Exactly how this gets explained will, no doubt, depend on background metaphysical assumptions. However, one option is to explain this fact in terms of locative relations. Each of the $P_n$ is located in a different place, and none of those locations are close enough to each other for a relation of spatiotemporal contiguity to obtain.

We can roughly formulate this *location pattern* as a general argument pattern as follows:

(1) For any objects $P_1 \ldots P_n$, if the $P_1 \ldots P_n$ do not stand in location relations $L_1 \ldots L_n$ to regions that stand in relations $R_1 \ldots R_n$, then the $P_1 \ldots P_n$ do not stand in relations $R_1 \ldots R_n$.

(2) The $P_1 \ldots P_n$ do not stand in location relations $L_1 \ldots L_n$ to regions that stand in relations $R_1 \ldots R_n$.

Therefore,

(3) The $P_1 \ldots P_n$ do not stand in relations $R_1 \ldots R_n$.

The filling instructions tell us that the $P_n$ are concrete objects, the $R_n$ are a class of relations in which regions or concrete objects might stand, and the $L_n$ are locative relations. The classification tells us that (3) follows from (1) and (2) by modus ponens.

The thought is that spatiotemporal contiguity is one of the relations that the $P_n$ fail to instantiate in virtue of their locations. Spatiotemporal contiguity is not the only relation that the $P_n$ might fail to stand in, however, in virtue of their locations. They might fail to be arranged from largest to smallest, despite being spatiotemporally contiguous. Or they might fail to be stacked on top of one another despite being spatiotemporally contiguous and so on.

The location pattern is superior to the problematic modus tollens composition pattern. The reason for this is that the modus tollens version of the composition pattern can be used to explain the failure of objects to stand in a relation of spatiotemporal contiguity only. The location pattern, however, could be used to explain the failure of objects to stand in other kinds of relations as well, depending on their locations (such as those just specified). Given the choice, then, we should select the location pattern to be a part of the generating set rather than the modus tollens composition pattern. This means that we can allow the modus ponens composition pattern entry into the generating set without also being forced to accept the modus tollens composition pattern. So, we can ensure that the intuitive direction of explanation in this case is respected.

We suspect a similar solution to be available for each putative case of metaphysical explanation. One of the features of metaphysical explanation appears to be that facts at lower levels explain many more facts at higher levels: metaphysical explanations move 'up' what Wilsch (2015/2016) calls the axis of fundamentality. So, for instance, location relations at one level explain relations of spatiotemporal contiguity at a higher level, plus various arrangement relations between objects (stacking, ordering and so on). Thus, by looking 'up' the ladder of metaphysical explanation we can infer a great deal about higher levels from lower level location relations. If we look 'down' the ladder of metaphysical explanation, by contrast, then it is difficult to see what we can infer about location relations, if anything. We don't get anywhere near as much inferential bang for our buck.

In short, there are good reasons to suppose that a generating set including argument patterns that move 'up' the ladder of metaphysical explanation will satisfy the criterion of paucity to a greater degree than a generating set which makes use of patterns that look 'down' the ladder of metaphysical explanation. Of course, a generating set including downward-looking patterns might mitigate this loss of paucity by a corresponding gain in stringency. However, it is doubtful that these downward looking-patterns will be very stringent. To see why, imagine that this ladder of metaphysical explanation has a forking structure. The nodes in the fork are entities, and the fork itself is a relation of metaphysical explanation. The ladder goes from fewer entities at the lowest levels up to a multitude of entities at the highest levels. Because of this forking structure, when we look down the ladder, we will need to rope together facts about a number of disparate entities at a given rung in order to be able to derive even a single fact about an entity at a rung that is lower down. While it is no doubt possible in many cases to build such a pattern,[12] the derivations that are instances of that pattern will display a high degree of heterogeneity. Compare this to the patterns that model the upwards direction of the ladder. In order to derive a single fact about a higher level entity, we typically do not need to rope together a disparate group of facts about lower level entities. So, the patterns that model the

[12] Though not in all cases, as information is lost as we progress up the ladder. For instance, it is clear how we can derive that objects have determinable properties on the basis of their determinate properties, but less clear how we can derive that an object has particular determinate properties on the basis of the determinable properties it instantiates.

metaphysical ladder going up should be more stringent than the patterns that model the metaphysical ladder going down.

Now, when we are faced with a modus ponens argument pattern and a modus tollens version of the same pattern, these two different inferences take us in different directions along the ladder of metaphysical explanation. Assuming that facts about lower level entities typically imply more at higher levels rather than vice versa, we can expect that only one of these patterns will be in the generating set. The one that will be in the generating set will be whichever pattern is following the ladder of metaphysical explanation upward. In other words, $E(K_M)$ ought to track the direction of metaphysical explanation in exactly the right way.

What if there is no ladder of metaphysical explanation of the kind that we have just described? What happens if there is no inferential asymmetry between facts at lower levels and facts at higher levels? In this situation, it may be quite difficult to get the unificationist machinery to line up with the intuitive metaphysical explanations. A modus tollens pattern may be just as good as the associated modus ponens pattern. Because we find the assumption at issue plausible—that facts about lower level entities explain more at higher levels—we are willing to make the unificationist theory a hostage to fortune.

## 5 Conclusion

The unificationist theory successfully rules that reflexive, monotonic, symmetrical and irrelevant derivations are not metaphysical explanations. It is thus a viable contender for an account of metaphysical explanation. There is, of course, work to be done. We need to show that the theory is not just viable but *adequate*. Doing that requires showing that the best systematisation of our metaphysical beliefs lines up with our intuitions about what is, and what is not, a metaphysical explanation. Further testing of the unificationist theory by looking at more particular cases of metaphysical explanation and fitting them to argument schemas is needed. Nonetheless, we hope to have provided the necessary motivation to undertake such a task.

## References

Audi, P. (2012). A clarification and defense of the notion of grounding. In F. Correia & B. Schnieder (Eds.), *Metaphysical grounding: Understanding the structure of reality*. Cambridge: Cambridge University Press.

Daly, C. (2012). Skepticism about grounding. In F. Correia & B. Schnieder (Eds.), *Metaphysical grounding: Understanding the structure of reality*. Cambridge: Cambridge University Press.

Dasgupta, S. (2014). On the plurality of grounds. *Philosophers' Imprint, 14*(20), 1–28.

Dasgupta, S. (2017). Constitutive explanation. *Philosophical Issues, 27*(1), 74–97.

Fine, K. (2012). Guide to ground. In F. Correia & B. Schnieder (Eds.), *Metaphysical grounding: Understanding the structure of reality*. Cambridge: Cambridge University Press.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48,* 507–531.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation*. Minneapolis: University of Minnesota Press.

Norton, J., & Miller, K. (2017). A psychologistic theory of metaphysical explanation. *Synthese*. https://doi.org/10.1007/s11229-017-1566-x.

Raven, M. J. (2012). In defence of ground. *Australasian Journal of Philosophy, 90*(4), 687–701.

Raven, M. J. (2015). Ground. *Philosophy Compass, 10*(5), 322–333.

Rodriguez-Pereyra, G. (2005). Why truthmakers? In H. Beebee & J. Dodd (Eds.), *Truthmakers: The contemporary debate*. Oxford: Clarendon Press.

Schaffer, J. (2009). On what grounds what. In D. Manley, D. Chalmers, & R. Wasserman (Eds.), *Metametaphysics: New essays on the foundations of ontology*. Oxford: Oxford University Press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies, 173,* 49–100.

Shaheen, J. (2017). The causal metaphor account of metaphysical explanation. *Philosophical Studies, 174,* 553–578.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

Thompson, N. (2018). Irrealism about grounding. *Royal Institute of Philosophy Supplements*, *82*, 23–44.

Wilsch, T. (2015). The nomological account of ground. *Philosophical Studies, 172,* 3293–3312.

Wilsch, T. (2016). The deductive-nomological account of metaphysical explanation. *Australasian Journal of Philosophy, 94*(1), 1–23.

Wilson, J. (2014). No work for a theory of grounding. *Inquiry: An Interdisciplinary Journal of Philosophy, 57*(5–6), 535–579.

Wilson, A. (2017). Metaphysical causation. *Nous*. https://doi.org/10.1111/nous.12190.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.