A COUNTERFACTUAL APPROACH TO EXPLANATION IN MATHEMATICS

SAM BARON, MARK COLYVAN, AND DAVID RIPLEY

Department of Philosophy, University of Western Australia, Crawley 6009, Australia.
Department of Philosophy, University of Sydney, Camperdown, 2006, Australia.
Department of Philosophy, Monash University, Caulfield 3162, Australia.

1. Unification

Call the explanation of one mathematical fact by another an *intra-mathematical explanation*. To date, there has been a tendency to approach the topic of intra-mathematical explanation by investigating the distinction between explanatory and non-explanatory proofs (see, for instance, [14; 22; 31]). This is very natural since it is widely acknowledged that some proofs are explanatory while others are not [16, p. 879]. Still, focussing exclusively on proofs as the only locus of explanation in mathematics is a mistake [15; 24]. That would be to prejudice the question of where explanations in mathematics are to be found.[1]

As with other cases of explanation, we should be asking "what makes a particular explanation explanatory?" Jumping to the question "which proofs are explanatory?" introduces a restriction on the theoretical options available for understanding intra-mathematical explanation. For the time being, at least, we'd like to remain open minded about where explanations in mathematics reside.

Ultimately, then, what we seek is a theory of intra-mathematical explanation that is capable of telling us how such explanations work, one that avoids restricting itself from the outset to asking only after proofs. In this paper, we explore a counterfactual approach. However, we will not be offering a full-blown counterfactual theory of intra-mathematical explanation just yet. Instead, we will offer a preliminary theory and show that the explanatory structure of intra-mathematical explanations can be modelled using counterfactuals. This clears the way for the

We are indebted to a referee of this journal for pressing us on many points and for several helpful suggestions. Material from this paper was presented in the Berkeley Logic Colloquium at the University of California, Berkeley and the Munich Centre for Mathematical Philosophy at the Ludwig-Maximilians University, Munich. Sam Baron's work on this paper was supported by an Australian Research Council Discovery Early Career Research Award (grant number: DE180100414). Mark Colyvan's work on this paper was supported by an Australian Research Council Future Fellowship (grant number: FT110100909) and a Carl Friedrich von Siemens Research Award of the Alexander von Humboldt Foundation.

Orcid.org/0000-0003-4000-3276. Email: samuel.baron@uwa.edu.au.

Email: mark.colyvan@sydney.edu.au.

Orcid.org/0000-0002-3356-0771. Email: davewripley@gmail.com.

[1] [43] makes this point in a more general way with respect to focusing too much on the aesthetics of proofs: [13] makes the suggestion that explanation in mathematics may arise in other places, such as domain extensions and reductions.

use of counterfactuals in coming to terms with intra-mathematical explanation and thus for the development of a full theory.

The prime motivation for our project is that of providing an account of explanation that will work wherever explanations arise. In addition to intra-mathematical explanations, there are at least two further varieties of explanation that must be taken account of. The first of these is extra-mathematical explanation: the explanation of a physical fact in part by mathematical facts (see [1; 2; 5; 11; 12; 21; 29; 30] for discussion). The second is physical explanation: the explanation of one physical fact by another (such explanations tend to be causal [41]; though perhaps not exclusively [39]). Counterfactuals have been appealed to in attempts to understand explanations of both kinds, most notably with respect to physical explanation [36; 47; 48] but also with respect to extra-mathematical explanation [2; 3; 4; 6]. By extending the use of counterfactuals to the intra-mathematical case, we therefore lay the groundwork for developing a unified account of explanation in science.

Although counterfactuals are closely related to explanation in science, one might resist the idea that counterfactuals enjoy a similar role within mathematics. One way to press this point is to ask for evidence from mathematical practice of such ties between counterfactuals and intra-mathematical explanation. Fortunately there is such evidence but a degree of caution is warranted here. One reason for caution is that mathematical explanation is not well understood so does not feature prominently in mathematics research papers. Talk of explanation features more in discussions of mathematics (by both mathematicians and philosophers) and in texts intended for teaching. Preliminary textual analysis suggest that there are many such cases, particularly in pedagogical texts aimed at student understanding.[2] Perhaps more important than the question of whether mathematicians *in fact* employ counterfactuals in the service of advancing explanation, is the normative question of whether mathematicians *ought* to employ counterfactuals in this way (or at all). This is harder to answer but a fruitful way forward is to see whether an appropriate account of counterfactuals in mathematics can shed light on mathematical explanation. The proof is in the pudding, as it were. In any case, this is the approach we take here. It is also worth noting that we are not suggesting that our particular counterfactual approach is the only way to shed light on mathematical explanation. We take this approach because it sits naturally with theories of explanation outside of mathematics. But there may be other ways to approach mathematical explanation; nothing we say here rules out such alternatives but we leave the exploration of such alternatives for another occasion. Our aim is to get one proposal suitably worked out and on the table for discussion.

We begin by briefly outlining a similarity-based approach to the truth of mathematical counterfactuals, along with a very general decision procedure for evaluating counterfactuals of this kind (§2). After that, we outline a particular case of mathematical explanation drawn from recent work by Marc Lange [22] and argue for three claims (§3). First, that there is counterfactual dependence of the explanandum on the explanans in this case. Second, that the counterfactual dependence within this particular case supports the asymmetry of explanation and, third, that the case can be situated within a structural-equation framework. We then turn

[2] See [38] for a pilot corpus analysis, which turned up a number of examples of counterfactuals and counterpossibles, many of these in the context of explaining the mathematical results in question and many instances from distinguished mathematicians (e.g. Terrence Tao [42, p. 311]).

our attention back to proofs and show that patterns of counterfactual dependence in the case of intra-mathematical explanation at issue reflect differences between explanatory and non-explanatory proofs (§4). Finally, we extend our discussion of counterfactual approaches to intra-mathematical explanation by considering a case involving impossible mathematics (§5).

2. Counterfactuals

We adopt a closeness-based account of the truth-conditions for counterfactuals. The standard closeness-based account provided by Lewis trivialises for mathematical counterfactuals [25; 40]. Accordingly, we will adopt an extension of that account that avoids triviality [7; 8; 10; 18; 19; 20; 28; 32; 33; 35; 37].[3] Such an account takes Lewis's closeness-based semantics and extends it across both possible and impossible worlds. The semantics can be stated as follows:

> **Analysis 1** $A \mathrel{\Box\!\!\rightarrow} B$ is true at a world ω iff some possible or impossible world in which both A and B are true is closer to ω than any possible or impossible world in which A is true and B is false, if there are any possible or impossible worlds in which A is true.

Following Lewis, we take the closeness of worlds to be a matter of similarity. When it comes to mathematical counterfactuals, we are interested in similarity with respect to mathematics. The closest worlds are, very roughy, the worlds in which the mathematical facts are the most similar to our own world. More carefully, we will focus on *intrinsic* similarity. Thus, the closest worlds are the worlds in which the intrinsic features of the actual mathematical facts are preserved. The extrinsic features, such as they are, do not make a difference to the closeness between worlds.[4] We will also assume Lewis's duplication-based notion of intrinsicality:

> Property P is intrinsic iff, for any two duplicate things, either both have P or neither does. [27, pp. 355–356]

[3] Of course, one might have some further reason for adopting a semantic theory that delivers the triviality of the counterfactuals at issue. For some recent arguments in this direction, see [46]; and for rebuttals to those arguments, see [9]. Arguments on both sides are reasonably well-rehearsed, and so we won't rehash them here.

[4] Cards on the table: the focus on intrinsic similarity has been reverse-engineered from the case we outline in the next section. Which is to say, we started from intuitions about which counterfactuals are true, and then worked our way back to an account of similarity that seems to deliver the right truth-values. Methodologically, we take this to be broadly in line with Lewis's approach to counterfactuals, when he writes that:

> The thing to do is not to start by deciding, once and for all, what we think about similarity of worlds, so that we can afterwards use these decisions to test [Analysis 1]. What that would test would be the combination of [Analysis 1] with a foolish denial of the shiftiness of similarity. Rather, we must use what we know about the truth and falsity of counterfactuals to see if we can find some sort of similarity relation — not necessarily the first one that springs to mind — that combines with [Analysis 1] to yield the proper truth conditions. It is this combination that can be tested against our knowledge of counterfactuals, not [Analysis 1] by itself. In looking for a combination that will stand up to the test, we must use what we know about counterfactuals to find out about the appropriate similarity relation — not the other way around. [26, p. 467]

One might worry about extending Lewis's similarity-based account of counterfactuals to mathematics. The problem, in a nutshell, is that any counterfactual variation to mathematics leads, inevitably, to inconsistency. This is a problem, one might think, because the impossible worlds that we must order via closeness with respect to the actual world will be inconsistent worlds. But such worlds, one might argue, are worlds in which everything and its negation is true: they are trivial worlds. We are willing to accept, as a potential outcome of our view, that the impossible worlds we must consider are inconsistent. We deny, however, that the worlds are trivial. Such worlds are not closed under a classical logic (which would lead to triviality) but, rather, are closed under some contradiction-tolerant logic (such as LP). With such a logic in hand, we can imbue the impossible worlds with a sufficiently rich structure so that there are all kinds of inconsistent mathematical structures (which are more or less inconsistent, depending on how many contradictions they feature). The truth of a mathematical counterfactual can then be understood in terms of the closest inconsistent structures when necessary: those that display a high degree of intrinsic similarity with actual mathematics, despite being inconsistent.[5]

We recognise that there are other semantic theories of counterfactuals available, and other ways to understand the similarity between worlds. On some other accounts, the particular case study that we develop in the next section may not work out in the manner that we suggest.[6] But we do not think that this would be a terminal problem for the counterfactual approach to mathematical explanation, though it may be a reason to doubt the particular version of that view developed in this paper. Even then, the application of the particular similarity-based account that we develop would still have value as a 'proof of concept' for the application of counterfactuals to mathematical explanations more generally. This proof of concept provides a useful starting point for the development of further counterfactual approaches in the future.

Having outlined an account of what makes mathematical counterfactuals true, we will now offer a method of evaluating such counterfactuals. The evaluation of a counterfactual typically proceeds via the following imaginative procedure. First, we hold fixed certain facts. The facts that are held fixed are taken to be invariant under counterfactual change. Second, we alter as many facts as we need to in order to make the antecedent of a given counterfactual true. We call this 'twiddling'. Third, we carry the implications of a twiddle through the free facts — the facts we are not holding fixed. This is 'ramifying'. Ramifying typically proceeds via the facts that one is holding fixed. For one will typically be holding fixed very general principles of some kind — usually laws — which tell us what the implications of various changes are for the free facts.

[5] For examples of inconsistent mathematics — approaches to mathematics that are based on the use of inconsistent mathematical structures — see [34; 44; 45].

[6] In particular, it may be that a given counterfactual that appears to be intuitively true turns out to be false on the account we are offering. Whether that is a problem, depends on the status of the intuitions about the counterfactual, or about the case of explanation that the counterfactual is implicated in more generally. If the intuitions at issue are free-floating intuitions about what appears to be true or false, then perhaps we should reconsider the intuition in question in light of the theory. If, on the other hand, the intuition is something that is embedded in mathematical practice, then some modification to the counterfactual approach to explanation that we offer may be in order.

Obviously, when evaluating a counterfactual we don't hold fixed the consequent or the features that are mentioned in the consequent. But nor do we hold fixed facts that are 'downstream' from the consequent. What it means for a fact to be 'downstream' from the consequent depends on the counterfactual at issue. For most ordinary counterfactuals, the 'upstream' facts are the temporal facts that lie to the past of the antecedent, while the 'downstream' facts are the facts that lie to the future of the consequent. The reason that we don't hold fixed facts that are downstream of the consequent, is to allow space for the ramification procedure to work. The ramification procedure just is the process of working out the implications of a twiddle for these downstream facts. If we hold these facts fixed, however, then we cannot ramify the counterfactual in any non-trivial way.

To see this in action, suppose that one strikes a match and the match lights. Now consider the counterfactual: if the match had not been struck, it would not have lit. To evaluate this counterfactual, we hold fixed as much as we can about the intrinsic nature of the match and as much as we can about the laws of nature compatible with realising the antecedent. Thus, we hold fixed the laws governing combustion and friction to ensure that the counterfactual match behaves as much like the actual match as possible. We also hold fixed the fact that the match is made of wood, that the head is coated with phosphorous sulfide, that the match is a particular size and shape, since these are all intrinsic properties of the match. We exclude scenarios in which the match is made of, say, stone because they are not relevant to the counterfactual we are interested in, and are not necessary for realising the antecedent of the counterfactual. Holding fixed the relevant facts in this way, we then perform a twiddle: we make whatever changes we need in order to prevent the match from being struck. Finally, we carry the implications of the twiddle through facts that are downstream of the consequent — which, in this case, are facts about the future — via the physical laws of nature. If, post-ramification, the match lights anyway despite never having been struck, then the counterfactual is false. If not, then not.

The same broad picture applies to intra-mathematical counterfactuals. First, we hold certain facts fixed. As in the ordinary, physical case, the goal when holding fixed is to try to ensure a high degree of intrinsic similarity between the actual situation, and whatever counterfactual situations we end up considering.[7] In the mathematical case, as in the non-mathematical case, this means holding fixed as much as we can concerning the intrinsic properties of whatever mathematical features are mentioned in the antecedent of a given counterfactual, compatible with realising the antecedent itself. The less we hold fixed about the intrinsic properties of whatever we are interested in, the less confident we should be in the outcome of the evaluation procedure. That's because the counterfactual situation we end up considering may bear little resemblance to the actual scenario at issue in relevant respects (i.e., respects of intrinsic similarity).

We recognise no analogous presumption in favour of holding fixed extrinsic properties. This is so for two reasons. First, with respect to the features that are mentioned in the antecedent, at least some of the extrinsic properties will involve features that are mentioned in the consequent. If we hold fixed these extrinsic properties, then we may end up holding fixed the very features that the counterfactual

[7] In an important sense the decision about what to hold fixed is context sensitive — it's sensitive to the details of the counterfactual under consideration and the presumed contrast class.

aims to test, which will prevent the non-trivial ramification of the twiddle. Second, extrinsic properties are not important for similarity. Since the goal of holding fixed is to achieve a close match with respect to similarity, holding fixed the extrinsic properties in this way is unnecessary. Worse: doing so can skew the results of the evaluation procedure. A policy of generally holding fixed the extrinsic properties of the features mentioned in the antecedent of a counterfactual will restrict the ways of realising that antecedent to only those situations in which the extrinsic properties are present. This allows for features that are not relevant to the evaluation of the counterfactual to unduly influence the outcome of the evaluation procedure.

We also hold as many general mathematical principles fixed as we can (more on this in a moment). We do this for the same reason that we hold fixed the physical laws when assessing an ordinary counterfactual: we want the mathematical behaviour of the counterfactual situation to be as close to the mathematical behaviour of the actual situation as possible. Next, we 'twiddle' mathematics by making whatever changes we need to make in order to realise the truth of the antecedent. This may involve making some counterfactual change to mathematics, such as making a counterfactual change to a particular mathematical structure, or figure or to a broad mathematical principle. Third, we carry the implications of the change that we have made through the mathematical facts that are downstream of the consequent, via the general mathematical principles that we are holding fixed.

As noted, there is a more or less natural division between the upstream and the downstream facts in cases of ordinary counterfactuals. The division is due to the underlying temporal structure of the universe. The mathematical case lacks temporal structure, so it is less obvious what the relevant 'downstream' and 'upstream' facts might be, and thus it is perhaps less clear what we should hold fixed and what we should permit to vary in this case. But while there is no temporal structure, there is an analogous mathematical structure: nodes in the structure correspond to mathematical facts, and the links in the structure are asymmetric relations of mathematical dependence: the dependence of one mathematical fact on another. The facts that are 'upstream' from a given mathematical fact within such a structure, are the facts that the mathematical fact depends on. The facts that are 'downstream' from a given mathematical fact are the facts that depend upon that fact. When we carry forward the implications of a twiddle via the ramification procedure, we carry them through the mathematical facts that are downstream in this sense.[8]

Note that some of the downstream facts will be general mathematical principles. As noted, our recommendation is to hold fixed as many general principles as we can. We can now sharpen this up using the distinction between downstream and upstream facts. When evaluating a counterfactual with a mathematical antecedent, we hold fixed as many general mathematical principles that are upstream of the mathematical fact in the antecedent as possible. We don't hold fixed general mathematical principles that are downstream, so that the twiddle has the space to properly ramify.

Because mathematical facts are usually thought to be necessary, the 'twiddles' we are considering sometimes involve impossibilities. There are a number of objections one might raise against the idea of twiddling mathematics in this way. We will mention the three most common, and gesture toward a response in each case. This is well covered ground, so we shall be brief (for further discussion, see [6]).

First objection: since some mathematical twiddles are impossible, we cannot entertain the full range of alterations to mathematics at issue. However, mathematical impossibilities are no less conceivable than impossibilities more generally. Indeed, it is common to conceive of mathematical impossibilities, at least when they are not yet known to be impossible. For example, we can consider what things would be like if $P = NP$, and we can consider what things would be like if $P \neq NP$. One of these, however, is impossible; we simply do not know which. The situation for things known to be impossible is no different. Finding out that some mathematical statement is impossible might make some people *more reluctant* to conceive of what things would be like if it were true, but it does not make anyone *less able* to do so.

Second objection: when we twiddle the mathematics we end up changing the subject. For instance, suppose we consider what would have been the case, had 13 not been prime. Here's one way to do this: imagine that 13 had 2 and 6 as factors. But then, someone might respond, we're not talking about 13 anymore; we're talking about, say, 12, or 18. So we have failed to consider a situation in which 13 itself is not prime; instead, we've considered a situation in which '13' picks out some other number with 2 and 6 as factors.

[8] What is the relevant notion of mathematical dependence? It is tempting to think of it as provability. However, there are many cases in which two mathematical facts can be proved from one another, so provability does not give us a clean differentiation into 'upstream' and 'downstream' mathematical facts. One way forward is to accept that in such cases of mutual inter-provability there is explanation in both directions. Alternatively, one might supplement provability with a pragmatic constraint, such that only one direction of provability is singled out for a particular explanatory purpose, and for the evaluation of a particular counterfactual. In addition to provability, there are three further options. First, one might appeal to mathematical laws. Just as there are causal laws that impute particular asymmetric relations of causal dependence between facts — dependencies that can then be used to differentiate past from future facts when evaluating counterfactuals — so too might one think that there are mathematical facts that function as guiding principles in much the same way. Second, one might appeal to a more metaphysical notion of fundamentality, such as the recently-popularised grounding relation. One might say that the upstream facts are the ones that ground a given mathematical fact, and the downstream facts are the ones that are grounded in that fact. Finally, one might appeal to a notion of causation broad enough to apply to mathematical states of affairs, such as the one deployed by Zardini (see [49]). For present purposes, we leave the distinction between upstream and downstream facts at an intuitive level.

The same worry can be raised of any counterfactual. Suppose we imagine what would have been the case had Emmy Noether not been a mathematician. But then, someone might respond, we're not talking about Emmy Noether anymore; we're talking about, say, Mary Whiton Calkins, or Juliette Adam. So we have failed to consider a situation in which Emmy Noether was not a mathematician; instead, we've considered a situation in which 'Emmy Noether' picks out someone else. We take it to be clear that the objection has gone wrong as an objection to counterfactuals involving Emmy Noether not being a mathematician; we take it to be no less clear that the objection goes wrong as an objection to counterfactuals involving 13 not being prime. We can be sure that we're still talking about Emmy Noether because the person we are considering bears enough similarities to Emmy Noether (this is the strategy taken in [25]). Similarly, it can be the case that we're still talking about 13 because the number we are considering bears enough similarities to 13 (it comes after 12, before 14, has itself and 1 as factors, sums with 12 to make 25 and so on).

Third objection: when you make twiddles in mathematics you inevitably induce some contradiction so the prospects for sensibly evaluating counterfactuals involving mathematical facts are dim. There are two things to say here. First, it is doubtful that contradictions are inevitable. Indeed, in §3–4 we provide an example of counterfactual reasoning in mathematics that results in no contradictions whatsoever. Second, there is a handy method available for 'chasing away' contradictions. This is the same method used for ordinary counterfactuals. First, hold fixed in the manner that we have suggested. Next, make a twiddle. Finally, ramify through the facts that are downstream of the consequent. If a contradiction ensues, hold less fixed and retwiddle. If the ramification of the second twiddle also results in a contradiction, then hold less fixed and so on until the ramification can be carried out consistently. In short, to conduct the ramification, chase the contradictions out of the structure.

That being said, when 'chasing' contradictions, you must hold *enough* fixed to ensure that you are, in fact, evaluating the counterfactual you mean to be evaluating. So, for instance, consider again the counterfactual: if the match had not been struck, it would not have lit. When evaluating this counterfactual, we need to hold fixed the fact that there is a match. Given that we are holding this fact fixed, there are ways of preventing the match from being lit that we could consider that would result in a contradiction. For instance, suppose we hold fixed that there is a match and then consider a scenario in which the match doesn't light because it is disintegrated. Then we have a contradiction: there both is and is not a match. In this type of situation we should *not* stop holding fixed the fact that there is a match. Rather, we should consider a different kind of scenario in which the match fails to light. In short, we are not recommending a blanket policy to stop holding fixed whenever a contradiction is reached. This policy of holding less and less fixed should not threaten the basic facts needed to sensibly evaluate the counterfactual. This is true in non-mathematical cases, and so it should be extended to mathematical cases as well.

Even if we hold enough fixed to ensure that we are in fact evaluating the right counterfactual, one might still perceive a worry with this idea of 'chasing' contradictions. Suppose we give chase to contradictions that arise out of a twiddle on mathematics. Further, suppose that we chase the contradiction away by holding less and less fixed, but doing so results in a very different mathematical structure. In this situation, it may be that the counterfactual mathematical structure we are considering is just too far from any actual mathematical structure to give us much confidence in the outcome of the evaluation procedure. A counterfactual may be evaluated as true or false, but the truth-value is based on such a wildly different mathematical structure, that it is unclear what relevance that structure has for the actual mathematics we are considering.

There is, then, a potential tension in our evaluation procedure. On the one hand, we have recommended chasing contradictions out of a given mathematical structure when attempting to evaluate a mathematical counterfactual. On the other hand, we should not chase the contradictions out to the point that we have completely changed the subject. This tension becomes a problem when there is no way to ramify without inducing a contradiction somewhere. For in this situation, it may be that the only way to chase contradictions out is to consider a mathematical scenario that bears little resemblance to the actual scenario. What we recommend, then, is a limited chasing procedure. Instead of chasing the contradiction all the way through a mathematical structure, we chase the contradiction out of the neighbourhood of the consequent. What this means is that we push the contradiction far enough through the facts that are downstream of the consequent to be able to see whether the consequent is true or not given the twiddle that we have made.

This will, no doubt, involve leaving some contradictions in place, and so the evaluation procedure recommends that we consider inconsistent mathematical structures. What we are interested in, as noted, is the closest inconsistent structure: the one that is the most intrinsically similar to the actual mathematical structure. One might worry that reasoning about inconsistent mathematical structures is incoherent (quite apart from the metaphysical worry that the structures themselves are incoherent, which is the problem considered above). But, in fact, there has been a great deal of work in recent times to demonstrate the coherence of such reasoning (see [34; 44; 45]). To reason coherently about such structures, we recommend, as before, using a non-classical logic, that tolerates contradictions.

3. Intra-mathematical Explanation

Having outlined a strategy for evaluating counterfactuals within mathematics, we will now show how to model a particular case of intra-mathematical explanation using counterfactual machinery. First, however, it is important to say something about the relationship between counterfactuals and explanation. For present purposes, we will adopt a very basic account of this relationship, according to which explanation within mathematics is analysed in terms of counterfactual dependence. This can be captured by the following very rough counterfactual account of intra-mathematical explanation:

> **[The Basic Counterfactual Account]** A mathematical fact F explains another mathematical fact G iff $F \;\Box\!\!\rightarrow\; G$ and $\neg F \;\Box\!\!\rightarrow\; \neg G$.

The basic counterfactual account should be seen as a starting point for developing a more nuanced account of intra-mathematical explanation, rather than a final theory. In the service of developing this basic account, we will now do three things. First, we will show that the pattern of counterfactual dependence in a particular case of intra-mathematical explanation reflects the explanatory facts by showing that a counterfactual corresponding to the explanation in this case is true.

Second, we will show that the counterfactual dependence moves from *explanans* to *explanandum* and not vice versa in this case and thus that the asymmetry of explanation can be retained. Third, and finally, we will situate the case within a broader structural equation framework to bring the approach in line with counterfactual accounts of scientific explanation more generally.

Note that while we believe the counterfactual structure of the case that we consider reflects explanatory intuitions about it, we could be wrong and the counterfactuals may come apart from those intuitions. We don't see this as a deep problem. While we believe that our intuitions about what is and is not an explanation within mathematics should serve as a guide for developing a theory of mathematical explanation, it may be that not every intuition can be captured by any theory.[9] We should expect some intuitions to give way. When the counterfactuals come apart from our intuitions for a particular case, we should seek to apply the theory to a larger range of cases to test it. If the theory works in most cases, then perhaps we should just conclude that our intuitions are sometimes in error. In short, the development of a theory of mathematical explanation is a process of bringing our intuitions into reflective equilibrium with the theory, with an eye to maximising the fit between the theory and the target phenomenon delineated by the intuitions. The start of that process with respect to a counterfactual theory is the application of counterfactual modelling techniques to particular cases. Our main aim in what follows is to show that at least some intuitions about explanatory dependencies in mathematics can be captured by an appropriate counterfactual structure. We will take the case in question to offer a genuine explanation but, again, nothing hangs on this.

3.1. **Counterfactual Dependence.**

The case that we will focus on is drawn from a recent book by Marc Lange [22]. Suppose that ABCD is an isosceles trapezoid such that AB is parallel to CD, AD = BC, AM = BK and ND = LC (see Figure 1). Then, in ABCD, $|ML| = |KN|$.

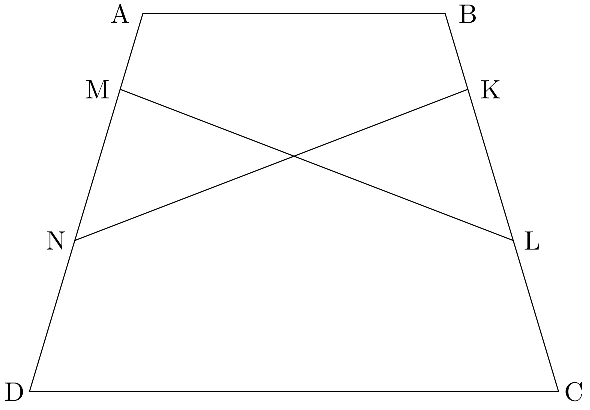

Figure 1: Isosceles Trapezoid ABCD

Why does $|ML| = |KN|$? The answer to this 'why' question lies with the symmetry of the trapezoid at issue. The isosceles trapezoid ABCD is really the same figure twice over, reflected along a line of symmetry that bisects the two bases (see Figure 2). Because ABCD is reflected along this line of symmetry the

[9] Indeed, there may be disagreement about the intuitions in question.

line segment MO is the same length as the line segment KO, and the line segment NO is the same length as the line segment LO. Thus, MO + OL = KO + ON and so $|ML| = |KN|$. As Lange puts the point:

> The theorem (that ML = KN) "makes sense" in view of the figure's overall symmetry. Intuitively, a proof that fails to proceed from the figure's symmetry strikes us as failing to focus on "what's really going on": that we have here the same figure twice, once on each side of the line of symmetry. Folding the figure along the line of symmetry, we find that NO coincides with LO and that MO coincides with KO, so that MO + OL = KO + ON, and hence ML = KN. ([23, pp. 246–247])

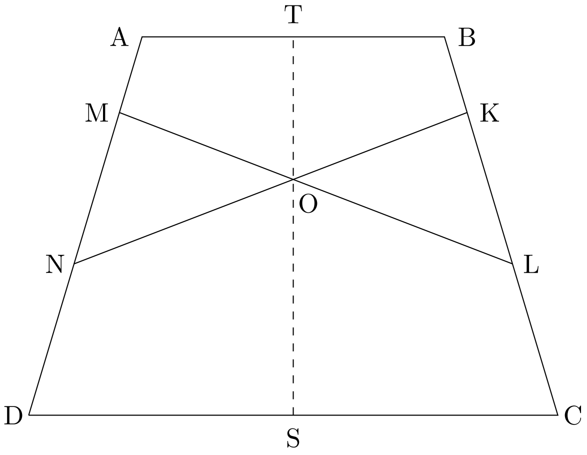

Figure 2: The Line of Symmetry in ABCD

We will return to proofs later on. For now we wish to focus only on the explanatory structure of the situation: the reason why $|ML| = |KN|$. As already noted, $|ML| = |KN|$ because ABCD is symmetrical. We suggest that this explanation is reflected in the counterfactual structure of the case. Consider the following open shape: α. α is made up of all of the line segments that constitute the left hand side of ABCD across the line of symmetry. This includes the line segment AD, the line segment AT, the line segment DS, the line segment from M to the line of symmetry ST, and the line segment NO to the line of symmetry (see Figure 3).

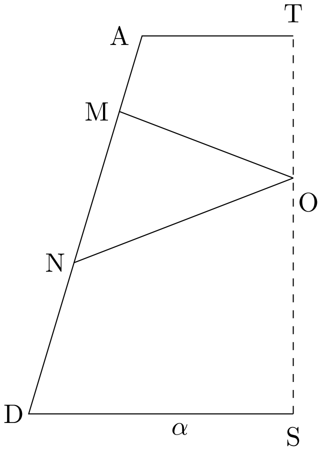

Figure 3: The Open Shape α.

Now, consider the following counterfactual about the isosceles trapezoid ABCD:

> CF$_1$ If, in ABCD, the open shape α had not been horizontally reflected, then it would not have been the case that $|ML| = |KN|$.

In order to evaluate this counterfactual, we use the method of evaluation described above. We begin by holding fixed as much as we can about the intrinsic properties of any features mentioned in the antecedent of the counterfactual, compatible with twiddling the antecedent.[10] There are two such features mentioned: the trapezoid ABCD and the open shape α. With respect to α, we should aim to hold as many of the intrinsic properties of that open shape fixed as we can. Since duplicating α would mean duplicating all of the line segments that make up that open shape, we must hold fixed the position, length, size and angle of every line segment featured. In short, all of α should be held fixed to ensure the highest degree of similarity.

Indeed, given that the counterfactual is centrally about α, changing the intrinsic properties of α would be like, in the match case, changing the intrinsic properties of the match when we consider a scenario in which it is not struck. While we can certainly consider such a scenario in which the match is made from stone, that scenario is not similar enough to the actual scenario and the actual match to give us any confidence in the outcome of the evaluation procedure. Similarly, in the present case, if we allow the intrinsic nature of the open shape α to change across the counterfactual scenarios that we are considering, we should have less confidence in the outcome of the evaluation procedure. The highest degree of confidence is associated with the highest degree of match between the actual α and the counterfactual α. With respect to ABCD, by contrast, we cannot hold fixed everything about that shape while also realising the antecedent of CF$_1$. For suppose that we do, then there will not be any way to prevent the horizontal reflection of α in ABCD. Once we have relaxed some of the features of ABCD, however, it is then possible to consider scenarios in which α is not horizontally reflected. Of course, we still want to consider a shape that is as similar to ABCD as possible, compatible with realising the antecedent.

At first glance, it is not obvious how to realise the antecedent of CF$_1$ without also altering α. For example, suppose we prevent the horizontal reflection of α in ABCD by moving the vertex B as in Figure 4.

[10] As noted before, the decision about what to hold fixed will be guided by context here. In particular, we are working within Euclidean geometry so we must hold the underlying geometry fixed.

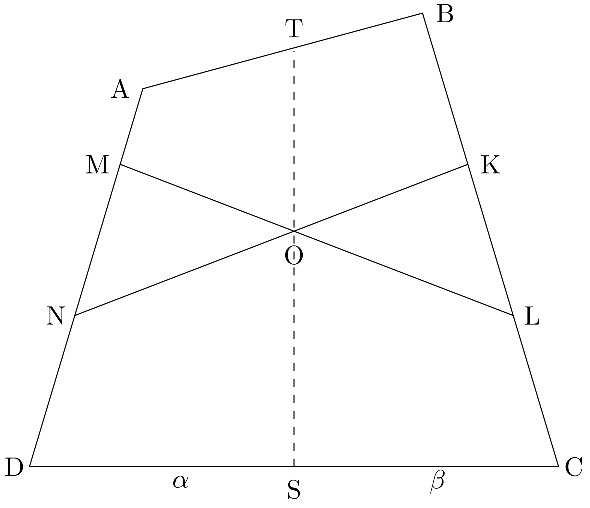

Figure 4: Moving B.

By moving B in this way, we end up changing α. The same thing happens if we move both B and C together along the x-direction, keeping their positions in the y-direction fixed (see Figure 5). In this case, we change α by altering the angles of the line segments MO and NO against AD.

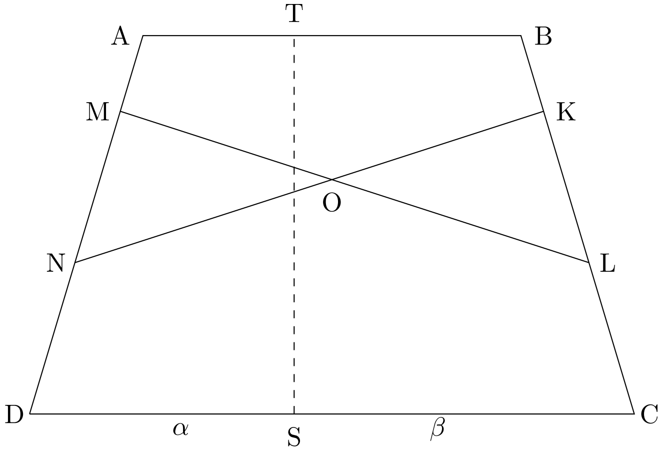

Figure 5: Moving BC along the x-direction.

The same result occurs if we leave B and C in place, and simply move K and L up and down the line segment BC (see Figure 6).

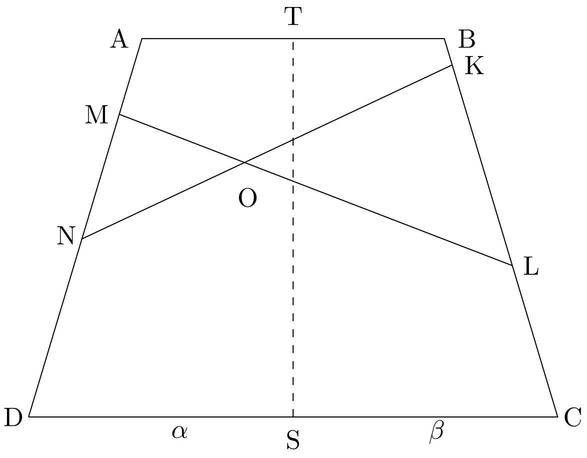

Figure 6: Moving K and L along BC

There is, however, a way to prevent the horizontal reflection of α in ABCD that fully preserves α. We must move BC along the x-direction *and* simultaneously move K and L along the line segment BC (see Figure 7).

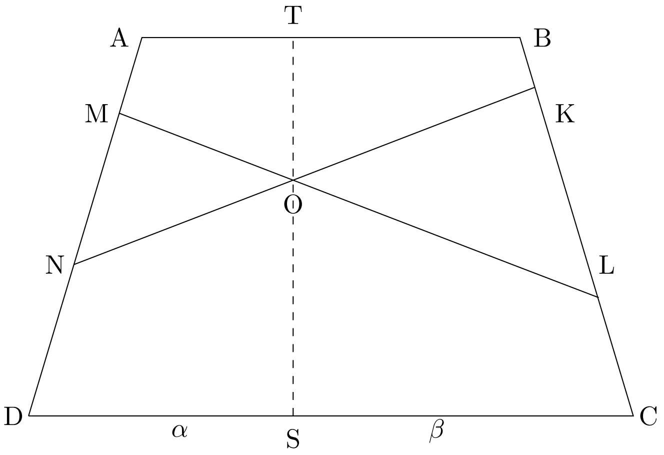

Figure 7: Moving BC along the x-direction, and moving K and L along BC as well.

Figure 7 is also the *only* way of preventing the horizontal reflection of α in ABCD while holding α itself fixed. There is no other way to make the antecedent of CF_1 true, so long as α is being held fixed. Having made this alteration to ABCD, we must now ramify that alteration through the rest of the figure. It is clear, however, that no matter how this change is implemented, the equality between $|ML|$ and $|NK|$ will be broken. That's because, in order to keep α fixed we have to ensure that the interior angles of △MNO remain unaltered when moving BC. But this means extending the line segments ML and KN to keep them in step with BC as we move it. Because of the angle of BC against the base CD, we will inevitably extend ML more than KN, thus making ML longer than KN (this is straightforward to prove algebraically). Because the only way of preventing the horizontal reflection of α in ABCD breaks the equality between $|ML|$ and $|NK|$, it follows that CF_1 is true: the only way of making the antecedent of that counterfactual true forces the consequent to be true as well.

3.2. **Asymmetry.** This completes the first stage of our modelling procedure. The next stage is to show that there is an asymmetry within the pattern of counterfactual dependence in this case that reflects the asymmetry of explanation. In order to show this, we need to consider a second counterfactual, namely:

> CF_2 If it had not been the case that $|ML| = |KN|$, then, in ABCD, the open shape α would not have been horizontally reflected.

In order to evaluate CF_2, we hold fixed as much as we can regarding the intrinsic properties of anything mentioned in the antecedent. Since the only features mentioned in the antecedent are the line segments ML and KN, we must hold the intrinsic properties of these two line segments fixed. Now, a line segment is just a very simple open shape. As with other open shapes, then, the key intrinsic properties of the line segment are its shape and size. Thus, an alteration to the line segment that turns it into a curve, would be a change to an intrinsic property. An alteration to the length of the line segment is another change in its intrinsic properties. Rotating and moving the line segment, however, is not a change to its intrinsic properties. The line segment is the same segment, whether it is oriented vertically, or horizontally so long as it links the same two points. This is because the intrinsic properties of shapes in general are preserved under rotation and translation. For instance, consider a simple circle. Rotating or moving the circle through a two-dimensional plane, does not alter the intrinsic nature of the circle. In order to duplicate the circle, we need only duplicate its shape and size. We don't need to also duplicate its relative location.

Notice, however, that it is not an intrinsic property of either ML or KN that the other line segment is present. For we can very easily imagine duplicating ML without duplicating KN or vice versa. All we need to do is duplicate the length and shape of the line segments in question. It follows that the intersection of the two line segments is not an intrinsic property of ML or KN either, since we can duplicate either line segment without the other and thus can duplicate both line segments without duplicating their intersection. So when evaluating CF_2 we don't hold fixed the fact that the two line segments intersect, or even the location and rotation of the two line segments. We also don't hold fixed α — the line segments that make up the left-hand side of ABCD — or ABCD more generally, since they fall into the consequent of the counterfactual.

Now, when evaluating CF_2, we can't hold fixed all of the intrinsic properties of the two line segments. In order to realise the antecedent, we need to break the equality between $|ML|$ and $|KN|$, which means changing the length of one or both of these line segments. Given this, the most straightforward way to realise the antecedent is by holding fixed the size and shape of one of the line segments, while altering the size of the other. This way of realising the antecedent results in a situation where α is not horizontally reflected in ABCD (see Figure 8).

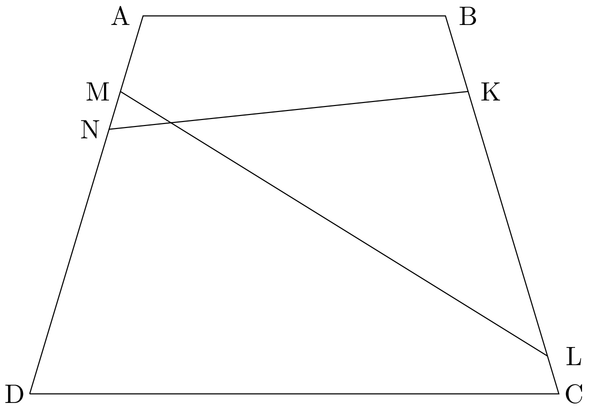

Figure 8: ML ≠ KN in ABCD which is not bilaterally symmetric.

If this were the only way of realising the antecedent of CF_2, then the counterfactual would be true. But there is another option. By rotating the two line segments, we can produce a situation in which one of the line segments is the same length as in actuality, while the other line segment is not (and thus $|MN| \neq |KL|$), but where α is horizontally reflected in ABCD (see Figure 9).

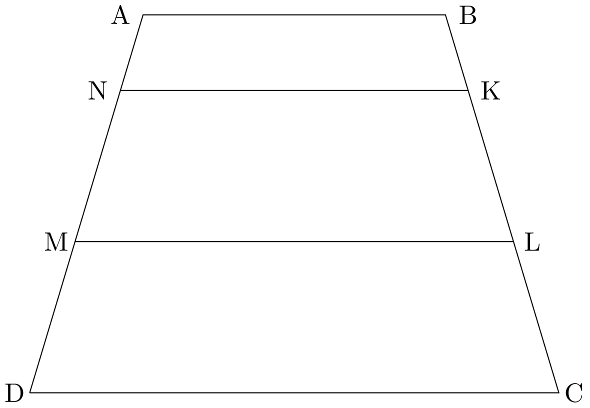

Figure 9: ML ≠ KN in ABCD which is bilaterally symmetrical; the length of ML is the same as it is in actuality.

These two ways of realising the antecedent of CF_2 involve changing the intrinsic properties of one of the line segments in the same way, while keeping the intrinsic properties of the other line segment fixed. So both options are equally similar to the actual situation, in relevant respects. Without a basis for privileging one of these options over the other (and, in particular, without a basis for preferring the symmetry-breaking cases to the symmetry-preserving cases) we cannot conclude that breaking the inequality between MN and KL *would* undermine the symmetry of ABCD. It *might* break the symmetry, but it might not, depending on exactly what we do. There is, then, an asymmetry between CF_1 and CF_2. For CF_1, there is really only one salient case in which we can make the antecedent true while maximising intrinsic similarity, and in that case the consequent is true as well. For CF_2, there are two classes of cases in which the antecedent is true that equally maximise intrinsic similarity, and no clear way to choose between them.

This asymmetry of counterfactual dependence, we submit, reflects the asymmetry of explanation.

3.3. **Structural Equation Modelling.**

In order to complete our modelling procedure we will show how to apply the standard tools of structural equation modelling to intra-mathematical explanation. We will briefly outline a structural equation model for the intra-mathematical explanation outlined above before dialling up the complexity of the model. First, however, it is important to address a general worry with the use of structural equations in this context.

Structural equations are themselves pieces of mathematics. So consider a particular intra-mathematical relationship that the structural equations are being used to encode. That relationship will, in line with the broad strategy for evaluating counterfactuals used above, involve twiddling some aspect of mathematics in order to see how that twiddle ramifies. All together the structural equations map the ramifications of the relevant twiddle. But what if the twiddle ramifies into the very mathematical facts that underwrite the structural equations themselves? Won't this undermine the entire structural equation framework?

The worry is based on a confusion. When we talk about 'twiddling' mathematics, we are not proposing to make the mathematical facts other than they are; we're not capable of doing any such thing. This is no different from non-mathematical cases; in evaluating "If Emmy Noether had not been a mathematician, she would have been a poet", we do not need to *make it the case* that Emmy Noether was not a mathematician! We are rather proposing to consider what *would be the case* if the mathematical facts were other than they are. But in our use of structural equation models, we are using *actual* mathematics, not counterfactual mathematics. It is no part of our proposal to consider what the structural equation models *would* tell us under the twiddles we consider. We are rather asking what these models *do* tell us, *about* those twiddles and how they ramify. So there is no problem; the structural equations themselves remain unchanged by the twiddles under consideration.

3.4. **A Basic Model.**

The model we will use to represent counterfactual relationships for the isosceles trapezoid ABCD is depicted in Figure 9. Each node in the diagram is a full proposition; each can be true or false. The initial model is designed to reflect the basic explanation for why it is that $|ML| = |KN|$ offered above. The point of the model in the first instance is to render a broad picture of the details of the case itself, and thus show how the modelling process works.

First of all, we provide a translation schema for the endogenous nodes. These are nodes that appear within the structural equation model and which represent the mathematical facts featured in the explanation. The translation schema is this:

$\mathcal{A}$	α is horizontally reflected.
$\mathcal{B}$	$\|ML\| = \|KN\|$.

With the translation schema in hand, we can then model the counterfactual relationships using the following directed graph.

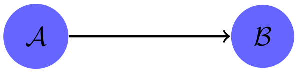

Figure 9: Directed graph model of the explanation for $|ML| = |KN|$.

Each node in the diagram takes value 1 or 0, according to whether the proposition it represents is true or false, respectively. We suppose that all the exogenous nodes take value 1. The structural equations we use are simple: each endogenous node takes the value(s) of the node(s) that feed into it. Thus, we have the following structural equations for the case at hand: $\mathcal{A} = 1, \mathcal{B} = \mathcal{A}$.

The structural-equation model captures the twiddles to mathematics involved in evaluating CF_1 in the previous section. Suppose that we twiddle node $\mathcal{A}$ by setting its value to 0. Then the value of $\mathcal{B}$ must also go to 0. The directed nature of the graph captures the asymmetry of the counterfactual dependence between $\mathcal{B}$ and $\mathcal{A}$ and thus between the fact that $|ML| = |KN|$ and the fact that α is reflected across ST in ABCD.

3.5. **Adding Complexity.**

The structural equation model that we have outlined is, obviously, quite simple. We can add complexity to the model in the same manner that we might add complexity to a simple structural equation model for a causal explanation. The most straightforward way to add complexity is to make the nodes non-binary. We can, for example, treat the nodes as representing continuous variables. Doing so allows us to take account of the relationship between explanans and explanandum with more sensitivity.[11]

As noted, we are interested in what happens to the relationship between the lengths ML and KN when asymmetry is introduced into the figure ABCD. As already discussed, the only way to make ABCD asymmetric whilst holding α fixed is to move the line segment BC along the x-direction, letting the vertices K and L move in step. So we can use the length between AB as a proxy for altering the figure ABCD in the required manner. As AB gets longer the difference between ML and KN should get larger. A structural equation model can then be used to capture the pattern of counterfactual dependence between, on the one hand, the difference between ML and KN and, on the other hand, the amount of extra length being added to the distance AB. We thus introduce two nodes:

$\mathcal{A}$	The length that is added to AB.
$\mathcal{B}$	The difference between the lengths ML and KN.

The structural equation can be determined as follows. First, it is useful to imagine a horizontal line segment from M to K parallel with AB, and another horizontal line segment from N to L, also parallel with AB. Call the point where the horizontal line segment from M intersects with the line of symmetry, ST, E and the point where the horizontal line segment from N intersects with ST, G (see Figure 10). Next, drop a perpendicular from B to the base CD. Finally, call the point where the perpendicular intersects the horizontal line segment from M, F, and call the point where the perpendicular intersects the horizontal line segment from N, H.

[11] It is common to treat the nodes in a structural equation model as representing the range of real values between 0 and 1. This is so that the nodes can be taken to represent probabilities. In his treatment of the structural equation framework, however, Pearl makes it clear that the values of the nodes can correspond to any quantity (e.g., length, see [36]). We follow Pearl in this.

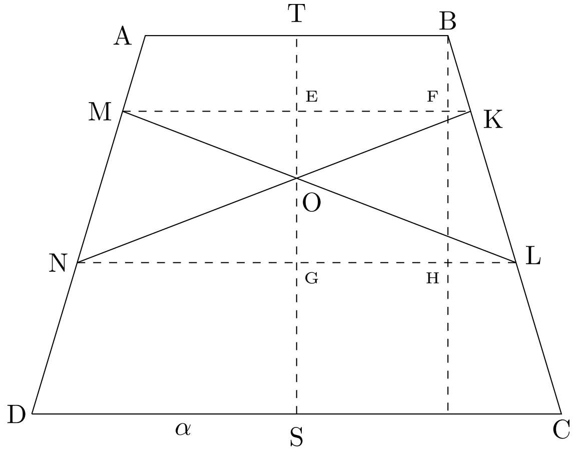

Figure 10: Setting up the Problem.

Call the difference between the lengths of ML and KN x and the amount of distance being added to the base AB y. The problem is to specify x as a function of y such that varying x leads to variations in y. Once specified, the function can then be used as the structural equation between the $\mathcal{A}$ and $\mathcal{B}$ nodes. Our strategy for solving the problem is to first determine each of the lengths of the line segments OK and OL in Figure 10 as a function of y. We can then use this information to specify x as a function of y. The problem can be solved as follows. Let us call the length of AB in the original figure: Δ and the distance between A and B in any transformed version of the figure ABCD $= \Delta + y$. First, specify $|OK|$ as a function of $\Delta + y$:

$|EK| = |EF| + |FK|$ [1]
$|EF| = |BT|$ because ETBF is a rectangle [2]
$|EK| = |BT| + |FK|$ [3]
$\frac{|FK|}{|BK|} = \sin(\measuredangle\text{ABC} - 90)$ [4]
$|FK| = (\sin(\measuredangle\text{ABC} - 90) \times BK)$ [5]
$|EK| = |BT| + \sin(\measuredangle\text{ABC} - 90) \times BK)$ [6]
$\frac{|EK|}{|OK|} = \cos(\measuredangle\text{OKE})$ [7]
$|OK| = \frac{|EK|}{\cos(\measuredangle\text{OKE})}$ [8]
$|OK| = \frac{|BT| + \sin(\measuredangle\text{ABC} - 90) \times BK)}{\cos(\measuredangle\text{OKE})}$ [9]
Since $|BT| = (0.5 \times \Delta) + y$, it follows that:

$|OK| = \frac{(0.5 \times \Delta) + y + \sin(\measuredangle\text{ABC} - 90) \times BK)}{\cos(\measuredangle\text{OKE})}$ [10]

Next, we specify the length of $|OL|$ as a function of $\Delta + y$. By parallel reasoning, it follows that:

$|OL| = \frac{(0.5 \times \Delta) + y + \sin(\measuredangle\text{ABC} - 90) \times BL)}{\cos(\measuredangle\text{OLG})}$ [11]
Now, since $|ML| = |MO| + |OL|$ and $|NK| = |NO| + |OK|$ and since $|ML| - |KN| = x$ it follows that:

$$x = (|MO| + \frac{(0.5\times\Delta)+y+\sin(\measuredangle\text{ABC}-90)\times BK)}{\cos(\measuredangle\text{OKE})}) - (|NO| + \frac{(0.5\times\Delta)+y+\sin(\measuredangle\text{ABC}-90)\times BL)}{\cos(\measuredangle\text{OLG})}) \; [12]$$

We can now swap x for $\mathcal{B}$ and y for $\mathcal{A}$ to yield the final structural equation, namely:

$$\mathcal{B} = (|MO| + \frac{(0.5\times\Delta)+\mathcal{A}+\sin(\measuredangle\text{ABC}-90)\times BK)}{\cos(\measuredangle\text{OKE})}) - (|NO| + \frac{(0.5\times\Delta)+\mathcal{A}+\sin(\measuredangle\text{ABC}-90)\times BL)}{\cos(\measuredangle\text{OLG})}) \; [13]$$

Finally, we want to remove the points E and G from the equation, so that the equation properly reflects the isosceles trapezoid ABCD depicted back in Figure 1. To do that, we exploit some of the symmetry of ABCD to redefine the angles ∡OKE and ∡OLF. This gives us two angles to deal with: γ and ϕ, which are defined as follows:

(1) $\gamma = 180 - (\measuredangle\text{NMO} + (180 - \measuredangle\text{MAB}))$
(2) $\phi = 180 - (\measuredangle\text{MNO} + \measuredangle\text{MAB})$

The reasoning behind the angle definitions is seen in Figure 11.

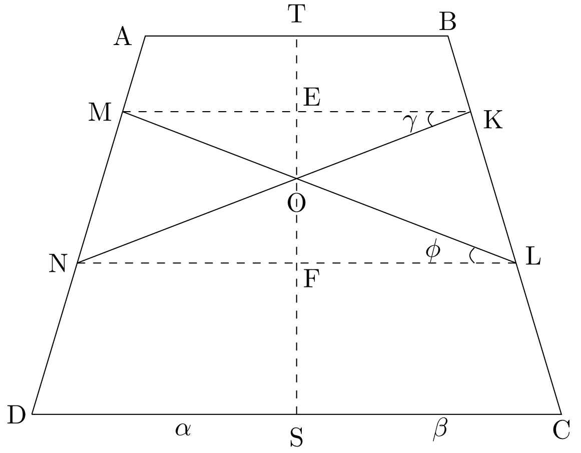

Figure 11: Because of the symmetry of ABCD, ∡NMO = ∡LKO. From the same symmetry plus the bisection of AB by MA and KB it follows that ∡EKB = (180 − ∡MAB). Thus, because NK bisects the straight line segment BC, $\gamma = 180-(\measuredangle\text{NMO}+(180-\measuredangle\text{MAB}))$. Similar considerations apply to ϕ. Because of the symmetry of ABCD, ∡MNO = ∡KLO. From the same symmetry it follows that ∡FLC = ∡MAB. Thus $\phi = 180 - (\measuredangle\text{NMO} + \measuredangle\text{MAB}))$.

Our structural equation, using the definitions just outlined, becomes:

$$\mathcal{B} = (|MO| + \frac{(0.5\times\Delta)+\mathcal{A}+\sin(\measuredangle\text{ABC}-90)\times BK)}{\cos(\gamma)}) - (|NO| + \frac{(0.5\times\Delta)+\mathcal{A}+\sin(\measuredangle\text{ABC}-90)\times BL)}{\cos(\phi)}) \; [14]$$

This final structural equation gets us what we want. When we change the value of $\mathcal{A}$ there will be a corresponding change to the value of $\mathcal{B}$. The change to $\mathcal{A}$ encoded in the structural equation is a change to the amount we are extending AB. The resulting change to $\mathcal{B}$ is the corresponding difference between the lengths of ML and KN that results from the twiddle to $\mathcal{A}$. The entire model then adequately captures the pattern of counterfactual dependence whereby gradual deformation of the object results in a widening inequality between ML and KN.

4. Explanatory and Non-Explanatory Proofs

In the previous section we showed how to model a case of intra-mathematical explanation using standard counterfactual machinery. In this section we will turn to the topic of explanatory versus non-explanatory proofs. Our contention is that the difference between explanatory and non-explanatory proofs can be captured by the patterns of counterfactual dependence underlying a given case of intra-mathematical explanation. To show this, we will consider three proofs of the fact that $|ML| = |KN|$ in the isosceles trapezoid ABCD. The first of these, given by Lange, is the proof discussed in §3, and is the one that Lange deems to be explanatory. He provides two alternative proofs that he takes to be not explanatory.

The first of these proceeds algebraically:[12]

> A proof could proceed by brute-force coordinate geometry: first let D's coordinates be (0,0), C's be (0,c), A's be (a,s) and B's be (b,s), and then solve algebraically for the two distances ML and KN, showing that they are equal. [23, p. 245]

The second proof is geometric, and proceeds as follows (see Figure 12):

> Draw the line from N perpendicular to CD; call their intersection P (see fig. 3); likewise draw link LS. Consider triangles DNP and CLS: angles D and C are congruent (since the trapezoid is isosceles), ND = LC (given), and the two right angles are congruent. Hence, by having two angles and the non-included side congruent, △DNP = △CLS, so their corresponding sides NP and LS are congruent. They are also parallel (being perpendicular to the same line). That these two opposite sides are both congruent and parallel shows PNLS to be a parallelogram. Hence, NL is parallel to DC. By the same argument with two new auxiliary lines, AB is parallel to MK. Therefore, MK and NL are parallel (since they are parallel to lines that are parallel to each other), so MKLN is a trapezoid. Since MN = AD − AM − ND, KL = BC − BK − LC, AM = BK, AD = BC and ND = LC, it follows that MN = KL. As corresponding angles, ∡KLN = ∡LCS; since △CLS = △DNP, ∡LCS = ∡NDP; as corresponding angles, ∡NDP = ∡MNL. Therefore, ∡KLN = ∡MNL. From this last (and that NL = NL, MN = KL), it follows (by having two sides and their included angle congruent) that △MNL = △KLN, and so their corresponding sides ML and KN are the same length. [23, p. 246]

[12] The C coordinate in Lange's geometric proof should be $C = (c, 0)$. As the editor for this journal, Robert Thomas, has pointed out to us, Lange's 'brute-force' proof contains an error. One must also specify that $b = c - a$, otherwise one has failed to specify an isosceles trapezoid. Rather than specifying a further condition, we ought to include this information in the specification of B as follows: let B be (c-a, s). From there, Thomas provides the proof as follows:

> M needs to be (ta, ts), N to be (ra, rs) where r and t are arbitrary parameters strictly between 0 and 1. Then for the needed equalities K is $[(1-t)c+t(c-a), ts]$ and L is $[(1-r)c + r(c-a), rs]$ from the standard formula for dividing a line segment in a ratio used twice. These complicated-looking co-ordinates simplify to $(c - ta, ts)$ and $(c - ra, rs)$ respectively. One can then just write down the distances squared from the distance formula, $|ML|^2 = (c-ra-ta)^2 + (rs-ts)^2$ and $|NK|^2 = (c - ta - ra)^2 + (ts - rs)^2$, obviously equal.

We are grateful to Thomas for providing us with the correct proof and note that Thomas prefers to treat it as 'straightforward' rather then 'brute-force'.

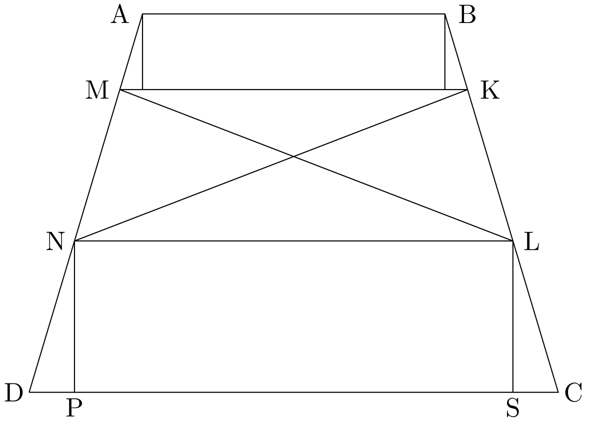

Figure 12: A Geometric Proof of $|ML| = |KN|$

Let us take each of these proofs in turn, starting with the algebraic proof.

4.1. **Algebraic Proof.** The central moving part of the algebraic proof is the coordinatisation of the isosceles trapezoid ABCD in a Cartesian plan. The algebra that follows (which we see no need to go into) is nothing more than a way of moving from the underlying coordinatisation to the fact that $|ML| = |KN|$. If we think of the matter in counterfactual terms, then for this proof to be a genuine explanation of the fact at issue at least CF_3 would need to be true (see below), since this counterfactual links the coordinatisation (the explanans) to the fact that $|ML| = |KN|$ (the explanandum):

> CF_3 If the coordinates of isosceles trapezoid ABCD had not been D = (0,0), C = (c,0), A = (a,s) and B = (c-a,s), then it would not have been the case that $|ML| = |KN|$.

In order to evaluate CF_3, we hold fixed all of the intrinsic facts about the trapezoid ABCD plus general mathematical principles and any mathematical facts that are upstream of the facts about ABCD. We then make a twiddle by re-coordinatising ABCD. There are many different ways to re-coordinatise the isosceles trapezoid ABCD that don't break the symmetry of the object and thus don't result in ML $\neq$ KN. We can, for example, reflect around the x and y axes (see Figure 13) or choose different coordinate systems. The symmetry of ABCD is insensitive to the particular Cartesian coordinate system used and there are infinitely many such coordinate systems.

Of course, there are also coordinatisations that will break the symmetry of ABCD. (E.g. non-linear translations of a Cartesian coordinate system such as: $x \mapsto x^2$; $y \mapsto y$.) There's an infinite number of those as well. In such coordinatisations, ML $\neq$ KN. What we have, then, is an infinite number of alternative coordinatisations in which ML = KN and an infinite number in which ML $\neq$ KN. We have no reason to prefer the symmetry preserving cases over the symmetry breaking cases when evaluating the counterfactual. The two cases are equally similar when it comes to the mathematical facts that we are holding fixed. It follows that CF_3 is false: it is not the case that an alteration to the coordinates of ABCD *would* result in ML $\neq$ KN. At best, it might have that result.

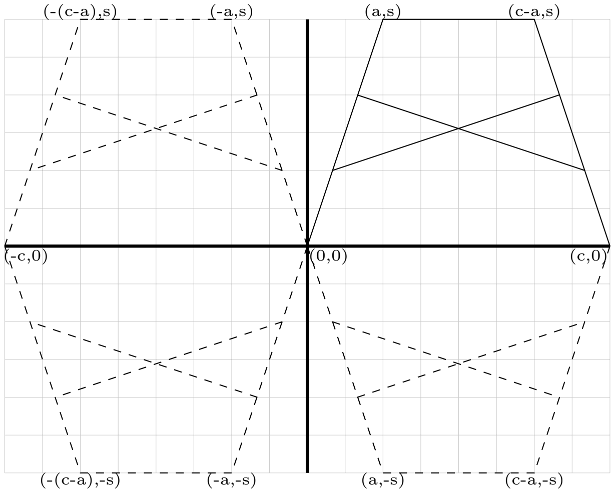

Figure 13: Symmetry preserving transformations of ABCD around the origin (0,0).

4.2. **Geometric Proof.** The first thing we need to do is clean up Lange's geometric proof.[13] As it stands the proof is inelegant. A more concise version of the proof can be constructed as follows. We begin by assuming that in a quadrilateral with a base and two equal opposite sides adjacent to it, the base angles are equal and the top is parallel to the base if and only if the quadrilateral is a trapezoid. It follows immediately that ∡ADC = ∡BCD because ABCD is a trapezoid. Next, draw the line NL. NLCD is a trapezoid and so it follows that NL is parallel to DC. NL is also parallel to AB (because DC is). ABLN is also a trapezoid, and so ∡MNL = ∡NLK. Since equals subtracted from equals are equal, $|AN| = |AD| - |ND| = |BC| - |LC| = |BL|$ and $|MN| = |AN| - |AM| = |BL| - |BK| = |KL|$ from which it follows that $|MN| = |KL|$. It follows that the △MNL and △NLK are congruent because $|MN| = |KL|$, NL is common between the two triangles and ∡MNL ≡ ∡NLK. Thus, $|ML| = |NK|$ since they are corresponding sides of △MNL and △NLK.

The proof hangs on the following facts:

(1) NL and AB are parallel
(2) NL and DC are parallel
(3) ∡MNL = ∡NLK
(4) $|MN| = |KL|$
(5) △MNL ≡ △KLN.

This gives us five counterfactuals to consider:

> CF$_4$ If it had not been the case that NL and AB are parallel then it would not have been the case that $|ML| = |KN|$.
>
> CF$_5$ If it had not been the case that NL and DC are parallel then it would not have been the case that $|ML| = |KN|$.
>
> CF$_6$ If it had not been the case that ∡KLN = ∡MNL then it would not have been the case that $|ML| = |KN|$.

[13] We are very grateful to Robert Thomas for supplying us with a version of the cleaned-up proof.

> CF$_7$ If it had not been the case that $|MN| = |KL|$, then it would not have been the case that $|ML| = |KN|$
>
> CF$_8$ If it had not been the case that $\triangle$MNL $\equiv$ $\triangle$KLN then it would not have been the case that $|ML| = |KN|$.

Notice that none of these counterfactuals mentions anything to do with the horizontal reflection of α. So we are under no obligation to hold fixed α when evaluating these counterfactuals. Indeed, because each of these counterfactuals involves altering α in some manner in order to make the antecedent true, we cannot hold α fixed without running into a contradiction.

CF$_4$–CF$_8$ are all false. Rather than going through all five counterfactuals one-by-one, we can gain a rough sense of why these five counterfactuals are false by comparing two pictures. The left-hand case in Figure 14 depicts a situation in which the antecedents of all five counterfactuals are true: NL is not parallel with AB or DC; $\measuredangle$KLN $\neq$ $\measuredangle$MNL; $|MN| \neq |KL|$ and $\triangle$MNL $\not\equiv$ $\triangle$KLN, and yet the consequents are all false because $|ML| = |KN|$ (this can be checked using the algebraic method discussed in §5.1). The right-hand case in Figure 14, by contrast, depicts a situation in which the antecedents of all five counterfactuals are true, and the consequents are also true because $|MN| \neq |KL|$.

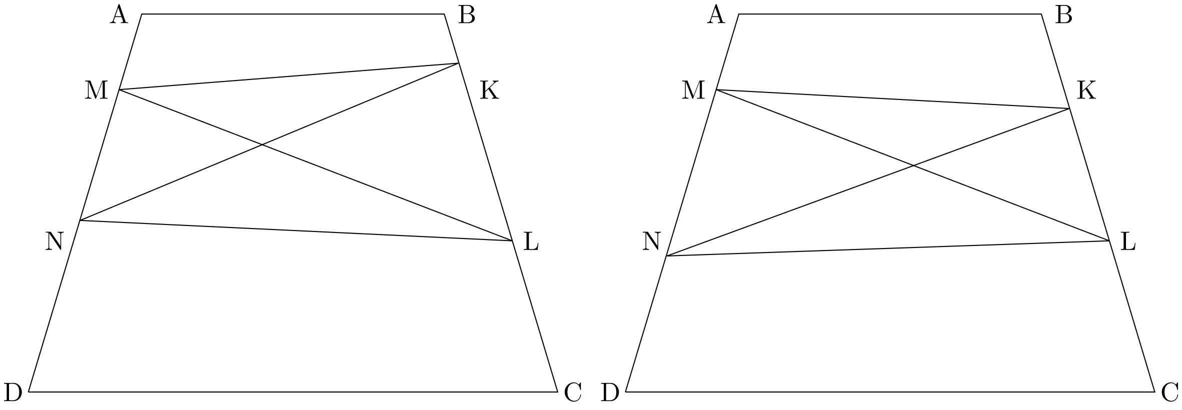

Figure 14: Situations that make CF$_4$–CF$_8$ true (LHS) and situations that make CF$_4$–CF$_8$ false (RHS).

The difficulty presented by Figure 14 should by now be familiar. In order to evaluate counterfactuals CF$_4$–CF$_8$ we hold the same things fixed in each case. In particular, we hold fixed as much as we can about the intrinsic properties mentioned in the antecedents, compatible with realising those antecedents. In addition, we hold fixed general mathematical principles and any mathematical facts that are upstream of the case at issue. We don't hold fixed the features that are mentioned in the consequent. There are two very similar ways to realise the antecedents, given what we are holding fixed. Figure 14 depicts two such cases. In the right-hand diagram, we have realised the antecedents in such a way that the counterfactuals all turn out to be false because $|ML| \neq |KN|$. In the left-hand diagram we have realised the antecedents of CF$_4$–CF$_8$ so that they all turn out to be true because $|ML| = |KN|$. It is difficult to visually discern the two cases, given how similar they are to one another. More importantly, these two cases are both equally similar to the actual situation, and so we have a tie. Without a way to break the tie, we cannot conclude that the counterfactuals are true.

4.3. **Explanatory versus Non-Explanatory Proofs.** It is striking that there is a difference in the pattern of counterfactual dependence between the proof of the fact that $|ML| = |KN|$ that Lange deems to be explanatory, and the proofs that he does not deem to be explanatory. The reason for this difference in the patterns seems to line up with Lange's account of why the explanatory proof in this case is explanatory, while the others are not. According to Lange, the explanatory proof is explanatory because it appeals to what is really doing the explanatory work, namely the symmetry of the figure ABCD. The other proofs do not appeal to this symmetry and so, in an explanatory sense at least, miss the point. Our own counterfactual analysis lends support to this idea. Apparently the only case in which we can get real traction on the explanation for $|ML| = |KN|$ using counterfactuals is when we look at alterations involving the horizontal reflection of α. When we look at other kinds of alterations — alterations that correspond to the facts used to derive $|ML| = |KN|$ in other proofs that ignore the symmetry of ABCD — none of these facts are counterfactually linked to the explanandum in the right way. Looking at the counterfactual structure of a given case of intra-mathematical explanation, then, may, at the very least, be a useful tool for sorting explanatory from non-explanatory proofs. It would also seem to follow that a counterfactual theory of intra-mathematical explanation would have a good chance of capturing the important distinction between explanatory and non-explanatory proofs more generally.

5. Impossible Twiddles

So far we have shown how to model a case of intra-mathematical explanation. One might worry, however, that we have made matters easy for ourselves by focusing on a purely geometric case. None of the counterfactual manipulations to this case that we have considered involves impossibilities in the ramification stage. After all, we are able to graph the various counterfactual alterations that are needed to model the intra-mathematical explanation for why it is that $|ML| = |KN|$ in the isosceles trapezoid ABCD (without drawing any impossible objects). We recognise, however, that at least some cases of intra-mathematical explanation will involve the consideration of impossibilities in the ramification stage. In this final section we will show how to extend the basic counterfactual machinery developed so far to one such case.

Consider the following mathematical fact: the product of any three consecutive, non-zero natural numbers is divisible by 6. The explanation for this fact appeals to two further facts. First, the fact that for any three consecutive nonzero natural numbers, at least one of those numbers is even and thus divisible by 2. Second, the fact that for any three consecutive nonzero natural numbers, exactly one is divisible by 3. From these two facts it follows that for any three consecutive numbers, their product is divisible by $3 \times 2 = 6$ [22, pp. 510–511].

As with our geometric case, the modelling of this basic number-theoretic case proceeds in three stages. First, we show that the explanandum counterfactually depends on the explanans. Then we demonstrate that the reverse is not true. Finally, we situate the example inside a structural equation model.

Because the explanation stated above rests on two salient facts, there are really two counterfactuals that correspond to the case. The two counterfactuals at issue are:

> CF_9 If it were not the case that for any three consecutive non-zero natural numbers, at least one of them is even, then it would not be the case that the product of any three consecutive, non-zero natural numbers is divisible by 6.
>
> CF_{10} If it were not the case that for any three consecutive non-zero natural numbers, at least one of them is divisible by 3, then it would not be the case that the product of any three consecutive, non-zero natural numbers is divisible by 6.

We can evaluate CF_9 as follows. First, hold fixed as much as we can about the intrinsic properties of numbers. Next, hold fixed as many general mathematical facts as we can, along with facts that are mathematically upstream from the numbers. This includes theorems about the natural numbers, so long as they are not downstream of the consequents of the above counterfactuals (more on this in a moment). Next, we twiddle the natural numbers by releasing only whatever mathematical theorems must be released to make it no longer the case that for any three consecutive non-zero natural numbers, at least one of them is even. The minimal way to do this is to pick three consecutive natural numbers and imagine that none of them is even. Finally, we consider the ramifications of this twiddle throughout the natural numbers in order to see whether or not the product of any three consecutive, non-zero natural numbers is divisible by 6.

By using this reasoning we can see that CF_9 is true. Consider the numbers 503, 504 and 505. $503 \times 504 \times 505 = 128,023,560$. The product of these three consecutive numbers is divisible by 6: $128,023,560/6 = 21,337,260$. Now, twiddle the natural numbers by making it so that none of 503, 504, or 505 is even. The twiddle ramifies as follows: the product of two natural numbers is even (if and) only if one of the numbers is. So $(503 \times 504) \times 505$ is even only if one of 503×504 or 505 is. Under the twiddle in question, 505 still isn't even, so we can turn to 503×504. By the same reasoning, this is even only if one of 503 or 504 is. Under the twiddle in question, neither of these is even. So 503×504 isn't even either, and thus $(503 \times 504) \times 505$ isn't even. Now, a number is divisible by 6 only if it is even; so $(503 \times 504) \times 505$ isn't divisible by 6. So under this twiddle, there are three consecutive numbers with a product that is not divisible by 6—so it is not the case that the product of any three consecutive, non-zero natural numbers is divisible by 6.

Similar reasoning renders CF_{10} true. Hold the same things fixed as when evaluating CF_9. Now, consider again the numbers 503, 504 and 505. This time, however, twiddle these numbers by making it so that none of 503, 504 or 505 is divisible by three. If none of 503, 504 or 505 is divisible by 3, then their product won't be divisible by 3 either. A number is divisible by 6 only if it is divisible by 3. So $(503 \times 504) \times 505$ isn't divisible by 6. So, once again, under the twiddle in question there are three consecutive numbers with a product that is not divisible by 6.

Now, one may worry that the counterfactual reasoning sketched above for CF_9 is a bit quick. If we continue to carry the ramifications of the twiddles in each case through the mathematical structure of the natural numbers, then we may well be forced to give up some fairly central number-theoretic theorems. For instance, consider the theorem that the sum of two even numbers is always an even number. If we make it such that 504 is not even, then this principle is called into question. After all, 500 and 4 are both even, and their sum is 504. If we hold fixed the evenness of 500 and 4, that their sum is 504, and the principle in question, we end up with a contradiction when we twiddle the evenness of 504. To avoid contradiction, then, we must not hold all of these fixed. This is just one example; the ramifications of making 504 odd may be great indeed. So great that one may well worry about the coherency of counterfactual reasoning of the kind under consideration. But we have already addressed this kind of worry. To conduct the ramification, chase the contradictions out of the immediate vicinity so that they may be ignored.

Demonstrating the truth of CF_9 and CF_{10} completes the first stage of modelling the case under consideration using counterfactuals. The second stage is to show that the following counterfactuals are false:

> CF_{11} If it were not the case that the product of any three consecutive, non-zero natural numbers is divisible by 6, it would not be the case that at least one of those three numbers is even.
>
> CF_{12} If it were not the case that the product of any three consecutive, non-zero natural numbers is divisible by 6, it would not be the case that at least one of those three numbers is divisible by 3.

Consider, first, CF_{11}. Suppose it were not the case that the product of any three consecutive, non-zero natural numbers is divisible by 6. It would not follow that at least one of those numbers is not even. Why? Because the fact that one of those three numbers is not even is not the only situation in which any three consecutive, non-zero natural numbers is not divisible by 6. As we have seen, if none of the numbers is divisible by 3, then the same result would follow. Symmetrical considerations apply to CF_{12}. CF_{12} is false because if it were not the case that the product of any three consecutive, non-zero natural numbers is divisible by 6, then it would not follow that none of those numbers is divisible by 3. For it may turn out instead that none of those numbers is even.

To sharpen the point it is useful to draw an analogy between the number-theoretic case under consideration and cases of *joint causation*. In a case of joint causation, there are two events E and E^* such that E and E^* are each necessary for the occurrence of an event E^{**}, and jointly (but not individually) sufficient. Such cases can be depicted as forking causal structures in which two events both contribute to the causation of a third event. The forking structure induces a counterfactual asymmetry: if E had not occurred then E^{**} would not have occurred, and if E^* had not occurred then E^{**} would not have occurred. But it would be wrong to say that if E^{**} had not occurred then E would not have occurred (because it may have been E^* that failed), and, symmetrically, it is wrong to say that if E^{**} had not occurred then E^* would not have occurred (because it may have been E that failed).

Horwich appeals to the fork asymmetry as a basis for the asymmetry of causation (see [17]). Our suggestion is that in cases of intra-mathematical explanation, similar fork asymmetries between mathematical facts underwrite counterfactual asymmetries of the kind demonstrated between CF_9 and CF_{10} on the one hand, and CF_{11} and CF_{12} on the other.

To draw this last point out a bit more, it is useful to move to the third stage of our modelling procedure: situating the case inside the structural equation modelling framework. The model we will use to represent counterfactual relationships in the number theory case is in Figure 15. To keep things simple, each node in the diagram is a full proposition; each can be true or false. The translation schema for the mathematical facts featured in the explanation is just this:

$\mathcal{A}$ For any three consecutive numbers, at least one of those numbers is even.

$\mathcal{B}$ For any three consecutive numbers, at least one of those numbers is divisible by 3.

$\mathcal{C}$ The product of any three consecutive non-zero natural numbers is divisible by 6.

Each node in the diagram takes value 1 or 0, according to whether the proposition it represents is true or false, respectively. We suppose that all the exogenous nodes take value 1. The structural equations we use are simple: each endogenous node takes the minimum of the values of the nodes that feed into it. This yields the following structural equations: $\mathcal{A} = 1$; $\mathcal{B} = 1$; $\mathcal{C} = \min(A, B)$.

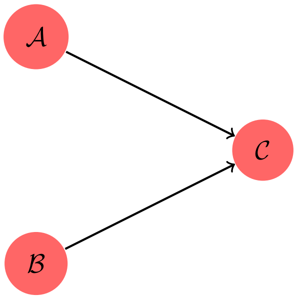

Figure 15: Directed graph model of CF$_9$ and CF$_{10}$

The structural-equation model captures the twiddles to the mathematics involved in evaluating CF$_9$ and CF$_{10}$ discussed in the previous section. Suppose we twiddle node $\mathcal{A}$ by setting its value to 0. The value of $\mathcal{C}$ will then be 0. Similarly, suppose we twiddle node $\mathcal{B}$ by setting its value to 0. Again, the value of $\mathcal{C}$ will be 0.

Because this is a directed graph, the asymmetry of the counterfactual relationships discussed above is encoded within the graph itself. Accordingly, setting $\mathcal{C}$'s value to 0 will not result in $\mathcal{A}$ or $\mathcal{B}$ taking the value 0, since that relationship is not reflected in the structural equations for the system. What we want to take note of, however, is the forking structure of the case. This forking structure is analogous to the forking structure of structural equation models of joint causation: cases where two events together contribute to the production of some event. It is interesting to draw out this parallel between the mathematical and causal cases, since it highlights a suggestive similarity between the explanatory structure of ordinary causal explanations and mathematical ones. This, in turn, speaks to the broad unificatory ambitions with which this paper began.

6. Concluding Remarks

In this paper, we've shown that there is good sense to be made of counterfactuals within mathematics and, moreover, that there is scope to apply our understanding of counterfactuals to cases of intra-mathematical explanation. The value of considering the application of counterfactuals to cases of intra-mathematical explanation lies with the project of developing a unified theory of explanation: a theory that can handle intra-mathematical explanation, extra-mathematical explanation, and physical explanation, treating all three as instances of a single phenomenon. No matter exactly how one conceives of that broader unificatory project, it is plausible that counterfactuals have some role to play, so by extending our understanding of counterfactuals to intra-mathematical cases, we have made progress toward the goal of unifying explanations within science.

References

[1] Alan Baker. Are there genuine mathematical explanations of physical phenomena? *Mind*, 114(454):223–238, 2005.

[2] Sam Baron. Optimization and mathematical explanation: Doing the Lévy walk. *Synthese*, 191(3):459–479, 2014.

[3] Sam Baron. The explanatory dispensability of idealizations. *Synthese*, 193(2):365–386, 2016.

[4] Sam Baron. Counterfactual Scheming. *Mind*, forthcoming.

[5] Sam Baron and Mark Colyvan. Time enough for explanation. *Journal of Philosophy*, 113(2):61–88, 2016.

[6] Sam Baron, Mark Colyvan, and David Ripley. How mathematics can make a difference. *Philosophers' Imprint*, 17(3):1–19, 2017.

[7] Jc Beall, Ross Brady, J. Michael Dunn, Allen Hazen, Edwin Mares, Robert Meyer, Graham Priest, Greg Restall, David Ripley, John Slaney, and Richard Sylvan. On the ternary relation and conditionality. *Journal of Philosophical Logic*, 41(3):595–612, 2012.

[8] Sara Bernstein. Omission impossible. *Philosophical Studies*, 173(10):2575–2589, 2016.

[9] Francesco Berto, Rohan French, Graham Priest, and David Ripley. Williamson on counterpossibles, *Journal of Philosophical Logic*, 47(4):693–713, 2018.

[10] Berit Brogaard and Joe Salerno. Remarks on counterpossibles. *Synthese*, 190(4):639–660, 2013.

[11] Mark Colyvan. *The Indispensability of Mathematics*. Oxford University Press, Oxford, 2001.

[12] Mark Colyvan. Mathematics and aesthetic considerations in science. *Mind*, 111(441):69–74, 2002.

[13] Mark Colyvan. *An Introduction to the Philosophy of Mathematics*. Cambridge University Press, Cambridge, 2012.

[14] Mark Colyvan, John Cusbert, and Kelvin McQueen. Two flavours of mathematical explanation. In Alexander Reutlinger and Juha Saatsi, editors, *Explanation Beyond Causation*, pages 231–249, Oxford University Press, Oxford, 2018.

[15] William D'Alessandro. Mathematical explanation beyond explanatory proof. *British Journal for the Philosophy of Science*, forthcoming.

[16] Timothy Gowers and Michael Neilson. Massively collaborative mathematics. *Nature*, 461:879–881, 2009.

[17] Paul Horwich. *Asymmetries in Time: Problems in the Philosophy of Science*. MIT Press, Cambridge, MA, 1987.

[18] Mark Jago. Impossible worlds. *Noûs*, 47(3):713–728, 2013.

[19] David Vander Laan. The ontology of impossible worlds. *Notre Dame Journal of Formal Logic*, 38(4):597–620, 1997.

[20] David Vander Laan. Counterpossibles and similarity. In Frank Jackson and Graham Priest, editors, *Lewisian Themes: The Philosophy of David K. Lewis*, pages 258–276. Oxford University Press, Oxford, 2004.

[21] Marc Lange. What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science*, 64(3):485–511, 2013.

[22] Marc Lange. Aspects of mathematical explanation: Symmetry, unity, and salience. *Philosophical Review*, 123(4):485–531, 2014.

[23] Marc Lange. *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford University Press, Oxford, 2017.

[24] Marc Lange. Mathematical explanations that are not proofs. *Erkenntnis*, 83(6):1285–1302, 2018.

[25] David Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[26] David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476, 1979.

[27] David Lewis. New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377, 1983.

[28] William G. Lycan. *Real Conditionals*. Oxford University Press, Oxford, 2001.

[29] Aidan Lyon. Mathematical explanations of empirical facts, and mathematical realism. *Australasian Journal of Philosophy*, 90(3):559–578, 2012.

[30] Aidan Lyon and Mark Colyvan. The explanatory power of phase spaces. *Philosophia Mathematica*, 16(2):227–243, 2008.

[31] Paolo Mancosu. Mathematical explanation: Problems and prospects. *Topoi*, 20(1):97–117, 2001.

[32] Edwin D. Mares. Who's afraid of impossible worlds? *Notre Dame Journal of Formal Logic*, 38(4):516–526, 1997.

[33] Edwin D. Mares and André Fuhrmann. A relevant theory of conditionals. *Journal of Philosophical Logic*, 24(6):645–665, 1995.

[34] Chris Mortensen. *Inconsistent Mathematics*. Kluwer Academic Publishers, Dordrecht, 1995.

[35] Daniel Nolan. Impossible worlds: A modest approach. *Notre Dame Journal of Formal Logic*, 38(4):535–572, 2001.

[36] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

[37] Graham Priest. *Beyond the Limits of Thought*. Oxford University Press, Oxford, 2002.

[38] Alexander Reutlinger, Mark Colyvan, and Karloina Krzyżanowska. The prospects for a monist theory of non-causal explanation in science and mathematics. to appear.

[39] Elliot Sober. Equilibrium explanation. *Philosophical Studies*, 43(2):201–210, 1983.

[40] Robert C. Stalnaker. A theory of conditionals. In Nicholas Rescher, editor, *Studies in Logical Theory*, pages 98–112. Blackwell, 1968.

[41] Michael Strevens. *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, Massachusetts, 2008.

[42] Terrence Tao. *Texts and Readings in Mathematics: Vol. 37. Analysis I.* Springer, 2016.

[43] R. S. D. Thomas. Beauty is not all there is to aesthetics in mathematics. *Philosophia Mathematica*, 25(1):116–127, 2017.

[44] Peter Verdée. Strong, universal, and provably non-trivial set theory by means of adaptive logic. *Logic Journal of the IGPL*, 21(1):108–125, 2013.

[45] Zach Weber. Transfinite cardinals in paraconsistent set theory. *Review of Symbolic Logic*, 5(2):269–293, 2012.

[46] Timothy Williamson. Counterpossibles. *Topoi*, 37(3): 357–368, 2018.

[47] James Woodward. *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press, Oxford, 2003.

[48] James Woodward and Christopher Hitchock. Explanatory generalizations, part 1: A counterfactual account. *Noûs*, 37(1):1–24, 2003.

[49] Elia Zardini. Instability and contraction. *Journal of Philosophical Logic*, 48(1):155–188, 2019.