Philos Stud (2021) 178:559–576
https://doi.org/10.1007/s11098-020-01445-8

Explanation impossible

Sam Baron[1,3] · **Mark Colyvan**[2]

Published online: 2 April 2020

Abstract We argue that explanations appealing to logical impossibilities are genuine explanations. Our defense is based on a certain picture of impossibility. Namely, that there are impossibilities and that the impossibilities have structure. Assuming this broad picture of impossibility we defend the genuineness of explanations that appeal to logical impossibilities against three objections. First, that such explanations are at odds with the perceived conceptual connection between explanation and counterfactual dependence. Second, that there are no genuinely contrastive why-questions that involve logical impossibilities and, third, that explanations appealing to logical impossibilities rule nothing out.

Keyword Explanation · Impossibility · Impossible worlds · Logic · Count

1 Introduction

Consider the following question and answer pair:

[Q] Why not *P*?
[A] Because *P* is impossible.

Call the answer here: an explanation by impossibility. There are different kinds of explanation by impossibility, depending on the modality at issue. For example, 'impossibility' might mean 'physical impossibility', in which case the explanation

✉ Sam Baron
samuel.baron@acu.edu.au

[1] University of Western Australia, Crawley, Australia

[2] University of Sydney, Camperdown, Australia

[3] Dianoia Institute of Philosophy, Australian Catholic University, 250 Victoria Parade, Fitzroy, VIC 3065, Australia

involves showing that the lack of *P* is guaranteed by the laws of nature. Or impossibility might mean legal impossibility, in which case the explanation appeals to the fact that *P* is forbidden by the law. And so on.

We are interested in explanations by impossibility, where the modality in question is logical impossibility. One place in which this issue has arisen recently concerns time travel scenarios. Suppose that Tim the time traveller travels back in time intent on murdering his grandfather before his mother was conceived. Tim spends his entire life training. He gathers documentation of his grandfather's movements in the past, and meticulously plans out the hit. In 2055, Tim travels back to 1985 to confront his grandfather as a young man. Tim shoots and fails. Grandfather goes about his day and Tim returns to the present.

It is plausible that Tim will fail in any such attempt. But why? According to Lewis (1976), Tim always fails for some *commonplace* reason. At the last moment, Tim's gun jams, or he has a sudden change of heart, or whatever. For Lewis, it doesn't really matter *what* prevents Tim from killing his grandfather. We simply know that there must be some such commonplace occurrence. We know this because we know that, as a matter of fact, Tim's grandfather was never murdered. This is evidenced by the very fact that Tim is able to travel back in time and make the attempt on his grandfather's life in the first place.

Some philosophers have expressed dissatisfaction with Lewis's proposed explanation. Arntzenius and Maudlin (2002, p. 180), Dowe (2007, p. 274), Gorovitz (1964, pp. 366–367), Horwich 1987, pp. 119–121), Riggs (1997, p. 52), Ismael (2003, p. 308) and Carroll (2010, p. 86) all worry that Lewis's proposed explanation leaves Tim's failed attempt on grandfather's life mysterious.

According to Baron and Colyvan (2016, 2019) there is a straightforward answer to this 'explanation' problem. One explanation for why Tim fails is by appeal to logical impossibility. Tim fails because it is impossible to succeed. Tim's success is forbidden by the laws of logic. After all, if Tim were to succeed, he would bring it about that he was never born. This, in turn, would bring it about that he never travelled back in time to kill his grandfather. So, if Tim succeeds, it follows that Tim also fails. Similarly, it follows that grandfather both lived because he avoided Tim's assassination attempt and died because he fell victim to it. Tim's success would violate the law of non-contradiction. This is not to say that the logical explanation is the only explanation in the neighbourhood. We can still appeal to Lewis's 'common place' occurrences to provide a low-level, causal explanation of Tim's failure. What the logical explanation provides is a high-level, non-causal explanation of the failure, one that explains why a common-place event must occur to prevent Tim's success.

Smith (1997, 2017) rejects this line of thought, arguing that the laws of logic play no role in explaining why Tim fails. Rather, according to Smith, the appeal to logical impossibility signals the *end* of explanation. Logical impossibilities never play a genuine role in explanation. At best, one can use logic to draw attention to the fact that there is no possibility of an explanation for *P* not being the case, or the fact that no explanation is required. Thus his view is that by pointing to the impossibility, one is indicating that there is something wrong with the 'why' question being asked. As he puts the point:

> the question itself is out of place ...it really should not be asked at all. There simply is nothing for him to succeed at — there is no such thing as a scenario that satisfies the description [of killing grandfather] and so there is no question as to why he fails to do 'that'. (Smith 2017, p. 166)

Smith goes on to suggest that *all* putative explanations that appeal to logical impossibility fail to be genuine explanations.

Our disagreement with Smith, although, about time travel, is of more general significance. The idea that logical impossibilities play an explanatory role is not just important for the grandfather paradox. Explanations of this kind also arise within logic itself. Consider, for instance, the following 'why' question: why is it the case that $P \rightarrow P$ for an arbitrary P? The answer proceeds as follows. Suppose that for some P, $P \rightarrow P$ were false, where '$\rightarrow$' signifies the material conditional. From this and the classical understanding of the material conditional, it would follow that $P\&\neg P$ is true. But that's impossible: it violates the law of non-contradiction. Thus, $P \rightarrow P$ is true because, at least in part, its falsity is impossible. In general, any logical proof that has the form of a *reductio* and that is explanatory is a case in which a logical impossibility seems to be playing an explanatory role. A similar thought can be extended to mathematics as well. Any explanatory proof in mathematics that uses a *reductio* step, seems to make use of a logical impossibility at some point.

Note that if it is logically impossible that P, then it is logically necessary that $\neg P$. One can just as easily appeal to the logical necessity of $\neg P$ to explain why $\neg P$ holds, as to the logical impossibility of P. We take it that explanations via logical necessity are just the flip-side of explanations by logical impossibility. Thus, any explanatory proof in logic that follows from logically necessary premises will count as a potential example of the kind of explanation that we have in mind.[1]

There hasn't been much written about the nature of logical explanation in general [though see Hoeltje et al. (2013), Schnieder (2008), Schnieder (2011), Schnieder (2016), Tsohatzidis (2015) for some work in this direction], and virtually nothing written about explanations that appeal to logical impossibilities. Our goal in this paper is to argue that alleged explanations appealing to logical impossibilities are, or at least can be, genuine explanations. Our defense is based on two presuppositions: that there are logical impossibilities and that the logical impossibilities have a certain structure (to be explained). Assuming this broad picture of impossibility, we defend explanation by appeal to logical impossibility against three objections. First, that such explanations are at odds with the perceived conceptual connection between explanation and counterfactual dependence. Second, that there are no genuinely contrastive why-questions that involve logical impossibilities and, third, that explanations appealing to logical impossibilities rule nothing out. We conclude

[1] Explanation by logical impossibility is closely related to the notion that there are impossible omissions that are causal and thus potentially explanatory. Bernstein (2016) defends this view, and some of the examples she uses could easily be turned into ones that use logical facts. Our thesis is also in line with the broad trend toward thinking in hyperintensional terms. See Nolan (2014) who considers explanation in this regard.

by defusing a further worry aimed at our presupposition that there are impossibilities.

2 Presuppositions

As noted, our argument is founded on two presuppositions. The first is simply that there are logical impossibilities. By making this assumption, we aim to stay completely neutral about the metaphysics of impossibilities. Our neutrality has two aspects. First, we are neutral about what impossibilities are, if they exist. A range of realist options are available in the wider literature on possible and impossible worlds for understanding the nature of impossibilities [see Mares (1997), Nolan (1997), Priest (2002), Restall (1997), Ripley (2012)]. Indeed, some have argued that any metaphysical account of possibility can be carried over to accommodate impossibility as well, at least if we are thinking of both possibility and impossibility in terms of worlds.[2] One can take these impossibilities to be sets of sentences in a world-building language, where the sets in question contain violations of the law of non-contradiction. One could adopt a realist position and hold that these impossibilities are concrete inconsistent worlds. And on it goes.

Our neutrality, however, is deeper than neutrality about what existing impossibilities are; we aim to be neutral about realism itself. Thus, by accepting that there are impossibilities in some sense, we are not thereby committing ourselves to a realist attitude about impossibilities. By saying that there are impossibilities, we mean only to accept that it is permissible to quantify over impossibilities. We allow, however, that one can take a fictionalist line or a broadly instrumentalist line with respect to this quantification and thus adopt an anti-realist attitude toward impossibilities. What matters for our purposes is that one can legitimately appeal to quantification over impossibilities in the context of providing an explanation.

The legitimacy of this appeal is thus not based on metaphysics. The idea is not that explanations appealing to logical impossibilities are genuine because impossibilities exist. Rather, the idea is that such explanations are genuine because one can reasonably appeal to impossibilities in order to address objections against such explanations.

Our second presupposition is that impossibilities have structure. A logically impossible situation does not imply that anything goes. The impossibilities are logically well-behaved. By this we mean, at a minimum, that the logical impossibilities are closed under a consequence relation of some suitable non-classical logic. There are many such logics to choose from, but for our purposes we will focus on the one that has received the most attention: the relevance logic *R*. *R* is weaker than classical logic. Thus, some *P* may be compatible with the laws of *R* and yet nonetheless involve a contradiction of some kind and so be incompatible with the classical laws of logic. Note that by 'compatible' here we just mean: the

[2] See Jago (2015) for an argument that not just any account of worlds will carry over. Even if Jago is correct, however, a number of viable options remain, such as the view defended by Ripley (2012).

conjunction of a contradiction and the laws doesn't lead to triviality (where triviality means that every proposition is both true and false).

We are willing to allow that the impossibilities have more structure than just the structure afforded by *R*. *R* implies less than classical logic, and so is a weaker logic in this sense. While some impossibilities are closed under *R*, there are logics still weaker than *R* that imply less, but that still imply something. We can thus impose a *weakness* ordering over impossibilities. One way to imagine this is using the familiar picture of concentric rings (an image we will return to later). The logical truths of classical logic, say, constitute a ring of necessity which limns the boundary of logical possibility. Everything within the ring is such that the logical truths of classical logic are all true. Beyond the ring of logical possibility we find a second ring: the ring of *R*, which is constituted by the logical truths of that logic. Everything within that ring is such that the logical truths of *R* are true. As we go outward, we see an expanding structure of concentric rings all the way out until we strike a logic in which *everything* is possible and *nothing* is necessary. All of the logical truths of this logic are true at every inner ring of this logic (there are no such truths, so the condition is trivially satisfied).

Our argument doesn't require this fuller structure, although such a structure is natural and appealing. It is, however, useful to have a picture to work with. Indeed, our argument doesn't require any particular account of the structure of impossibilities. We only need there to be some structure. It could be the logic *R* coupled with a weakness ordering. But equally, it could be another non-classical logic, and an ordering of a different kind. It may be that the picture of concentric rings is not apt for impossibilities, and what we have is a series of overlapping rings. Such a picture might invoke logical circles that are not nested within one another if, for instance, the sets of logical truths that define each ring do not stand in an iterative sub-set relation. Or it may be that the circles are not logical circles at all, but are produced in some other manner, perhaps by the introduction of a similarity ordering. Since similarity can be judged along different dimensions, the resulting picture may not involve anything like a nested circle conception. At any rate, impossibility is a strange place, and we make no specific claims about what to expect there. Just don't expect it to be a free-for-all, that's all we ask.

3 Three objections

In this section, we consider three objections to the idea that explanations appealing to logical impossibilities are genuine explanations. We will argue that all three objections fail.

3.1 Explanation and counterfactual dependence

As noted, the first objection focuses on the relationship between counterfactuals and explanation. For some, the relationship is one of conceptual analysis: explanation is defined in terms of patterns of counterfactual dependence (Lewis 1973a, 1979, 1986 comes close to this view). According to such a counterfactual account of

explanation (roughly) A explains B just in case: had A not been the case, B would not have been the case. For others, the relationship is weaker: explanations and counterfactual dependencies are mutually supportive, fitting together into a broader explanatory framework [see, for instance, Woodward (2003), Woodward and Hitchock (2003)]. At the very least, counterfactuals are an important tool for testing explanations.

The intimate relationship between counterfactuals and explanation is evident in standard explanations of why events occur, but it is also a feature of explanations by impossibility. For instance, consider a case of physical impossibility. Suppose one asks the following question: why does nothing with positive rest mass accelerate to 300,000 km/s? The answer appeals to the laws of nature: 299,792 km/s is the speed of light and acceleration past the speed of light is not possible, given the laws of relativistic mechanics. Corresponding to this explanation is the following counterfactual: if the speed of light had been 302,000 km/s, it would have been possible for an object with positive rest mass to accelerate to 300,000 km/s. This counterfactual is in good standing, in so far as any counterfactual is. It can, moreover, be used to work through the implications of the explanatory relationship between the speed of light as limiting speed and the acceleration of an object with positive rest mass.

The worry is that the connection between explanation and counterfactual dependence is severed in the case of explanations by logical impossibility. That's because counterfactuals involving logical impossibilities are not like counterfactuals involving physical impossibilities. One familiar way to press the worry is via the semantics for counterfactuals. This is well-covered ground and so we aim to be brief. On the standard Lewis-Stalnaker semantics for counterfactuals, the truth-conditions for counterfactuals are set by a closeness measure over possible worlds. Specifically, a counterfactual $A \mathrel{\square\!\!\rightarrow} B$ is true just in case all of the closest A-worlds are also B-worlds (Lewis 1973b; Stalnaker 1968). Given this account of the truth-conditions for counterfactuals, if A is logically impossible, then there are no A-worlds, so *a fortiori* every closest A-world is a B-world. We have triviality. The trouble, then, is that any counterfactual that might play a role in a case of explanation by logical impossibility will have a logically impossible antecedent. It will thus be trivially true by the Lewis-Stalnaker semantics. This triviality then threatens to infect the explanations themselves, or so the objection goes.

The solution here is straightforward. We note that the triviality result stems from the use of a particular semantic account of counterfactuals. There is, however, a natural extension of the Lewis-Stalnaker semantics available that avoids triviality. According to this extension, the truth-conditional analysis that Lewis provides for counterfactuals is basically correct; we just need to widen our conception of a 'world' by allowing the truth-conditions to range over both possible and impossible worlds (Bjerring 2014; Brogaard and Salerno 2013; Kment 2014). To make this modification explicit, we can rewrite the truth-conditions stated above as follows: $A \mathrel{\square\!\!\rightarrow} B$ is true just in case all of the closest possible or impossible A-worlds are B-worlds. This helps because while there may be no possible A-worlds, there are impossible A-worlds. With respect to those worlds, a counterfactual with an impossible antecedent may be true or false.

The triviality of counterfactuals involving logically impossibilities is thus by no means forced on us, at least not for semantic reasons. Some further argument in favour of triviality is needed. Some have attempted to provide such arguments [see, in particular, Williamson (2013)], but we are not convinced that such arguments succeed [see the responses in Berto et al. (2018); Baron et al. (2017)].

In the extended semantic picture, a closeness measure of impossibility is used; a measure that is analogous to the measure over possibility. One might worry that closeness is only meaningful for possibilities, but that doesn't seem very plausible. Closeness is just a measure of similarity. As noted, impossibilities are not an unstructured free-for-all. So, for example, a world exactly like the actual world except that there is one true contradiction is, by any reasonable lights, *very* much closer to the actual world than a world with different laws of physics, no people, and widespread contradictions.[3]

A more pressing worry might be that we haven't provided a viable similarity measure for the counterfactuals we are interested in. Lewis, for instance, provides a robust four-step recipe that tells us how to order worlds for the purposes of evaluating counterfactuals. We haven't offered anything like this. Agreed, but note that there are good reasons to suppose that Lewis's proposal doesn't work (though see Moss (2012) for discussion and defense of Lewis). In any case, it is not fully general: it doesn't work for backtracking counterfactuals [see Lewis's (1979) response to Fine (1975)], nor for counterfactuals that do not involve distinct events [see Lewis's (1987) response to Kim (1973)]. In fact, it is far from clear that there is a viable closeness ordering for any counterfactual [see Elga (2001) for criticism of Lewis's ordering, and Hájek (ms.) for criticism of a range of closeness accounts]. That's a reason to worry, in general, about the truth of counterfactuals (assuming a closeness semantics), but it does not seem to be a specific problem for counterfactuals involving logical impossibilities.

In so far as one yearns for a closeness ordering for the case of logical impossibility, we recommend heeding Lewis's advice about anything along these lines. According to Lewis, any such ordering must be reverse-engineered from our intuitions about the truth or falsity of counterfactuals. As he puts the point:

> 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 [the Lewis-Stalnaker semantics]. What that would test would be the combination of [the Lewis-Stalnaker semantics] 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 [the Lewis-Stalnaker semantics] to yield the proper truth conditions. It is this combination that can be tested against our knowledge of counterfactuals, not

[3] The idea that there is an impossible world that is closer than some possible world is ruled out by the so-called strangeness of impossibility condition (SIC). We are thus inclined to reject SIC. Even if one thinks it is true, however, it follows from this commitment that impossible worlds and possible worlds can be compared with respect to similarity (assuming closeness is a similarity measure).

> [the Lewis-Stalnaker semantics] 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 (Lewis 1979: [467])

As a number of philosophers have argued, we have robust semantic intuitions according to which some of these counterfactuals are true and some are false [see, for instance, Vander Laan (1997); Mares (1997); Nolan (1997); Restall (1997)]. So we recommend a similar process of reverse-engineering. The results, if they are any good, will be at least as good as Lewis's own recipe for counterfactuals.

One last outstanding issue. When dealing with counterfactuals involving logical impossibilities one may need to reason either with or about contradictions. So, for instance, one approach to the epistemology of counterfactuals appeals to imagination. We imagine a hypothetical scenario in which the antecedent holds, and then attempt to work through the implication of the antecedent, importing some but not all of what we know about the actual world. When the antecedent of a counterfactual is a logical impossibility, we are being asked to imagine contradictory scenarios and then reason coherently about them. It might be argued that reasoning with contradictions is something we just can't do, and so there is no viable epistemology to be had.

It is important to note that reasoning through a counterfactual with a logically impossible antecedent does not necessarily mean working through a contradiction. A logical impossibility is something that contradicts one or more logical truths. If we import these logical truths into the scenario we imagine then, yes, there will be some contradiction to consider. But that just means we shouldn't import every logical truth into such a scenario. This is a familiar point from reasoning about counterfactuals in general. If we import too much of what is actually the case into a situation we are imagining for any counterfactual, then we will get a contradiction (the most straightforward way to do this is to import the falsity of the antecedent).

One might think it is inevitable that some logical truth will get imported and, as a result, contradicted. But we don't see why the importation of logical truths should be inevitable. We are imagining that logical impossibilities are closed under a weaker logic such as R. So we don't need to import the actual logical truths into the scenario we are imagining to reason through it. But that's the only reason we can see for thinking that the importation of logical truths is mandated. Even if the importation of actual logical truths into the imagined scenario is mandated, there are logical truths of classical logic that aren't logical truths of R, and so there are some logical truths that we can import into R that won't result in a contradiction.[4]

Even if we have to reason about a contradiction, we have perfectly good frameworks for doing so. Logics such as R, and paraconsistent logics can be used to reason about contradictions and, indeed, have been used for precisely this purpose (in, for example, the study of naïve set theory). So while the epistemic process

[4] Indeed, whenever we engage in theory choice in logic we must at least entertain alternate logics — logics that disagree about what the logical truths are, for example.

might be unfamiliar, and might require a bit of training to be conducted successfully, it is possible.[5]

In sum, then, we admit that there is work to be done in understanding counterfactuals with logically impossible antecedents. But there is work to be done in understanding counterfactuals of any stripe. There is no special reason to doubt that counterfactuals with logically impossible antecedents are any worse in this respect. The ugly truth is that counterfactuals remain largely untamed, despite our best philosophical efforts. It is important to emphasise, however, that the same kinds of resources are available for understanding counterfactuals with logically impossible antecedents as for counterfactuals more generally. Just as we can say that there are possibilities and use them to understand counterfactuals with possible antecedents, we can say that there are impossibilities and we can use them to understand counterfactuals with impossible antecedents. In both cases, a semantic framework can be made to hang on the structure of possible and impossible situations. In both cases a closeness ordering can be produced that relies on degrees of similarity (though what the dimensions of similarity might be remains very much up in the air). And in both cases the closeness ordering can be reverse engineered from semantic intuitions, if need be.

3.2 Contrastive why-questions

We turn now to the second objection. This objection focuses on the nature of why-questions. It is plausible that all genuine why-questions are contrastive in nature. Thus, when we ask 'why P?' we should understand the request to be: 'why P and not Q?' for some contextually salient Q. Smith suggests that 'why not P?' isn't a genuine why-question, when $\neg P$ is logically impossible. One way to understand this worry is in terms of the contrastive nature of 'why' questions more generally: 'why not P' isn't a genuine why-question when $\neg P$ is logically impossible, because the why question fails to have a plausible contrastive form.

Here's the thought. Genuine why questions must be contrastive but they must also offer a contrast between two possibilities. The why-question 'why P and not $\neg P$?' fails to contrast two genuine possibilities when $\neg P$ is a logical impossibility. Instead, the why-question at issue offers a contrast between a possibility, and something that is not logically possible. Because why-questions involving logical impossibilities are not genuinely contrastive, *a fortiori* they are not genuine why-questions and thus explanations by logical impossibility are not genuine explanations. Compare this with cases in which $\neg P$ is a physical impossibility. In this situation the why-question 'why P and not $\neg P$' provides a contrast between two genuine possibilities.

Our response to this concern is to deny that why-questions must be contrastive between two possibilities. To be sure, why-questions must provide a contrast between two situations, but some of those situations can be impossible situations, as

[5] There is a further worry about whether we can even imagine impossible situations. We address this issue in the next section.

is the case with why-questions involving logical impossibility. After all, there is no blanket ban on impossibilities quite generally from serving as the contrast in a meaningful why-question. In cases of physical possibility, there are why-questions that contrast physical possibilities with physical impossibilities. This is perfectly coherent, so long as there are impossibilities of the relevant kind to serve in the contrast. And so it is in the logical case, if there are logical impossibilities, then they can constitute the contrast class for why-questions that arise in cases of explanation by logical impossibility.

Perhaps there is something about the relevant impossibilities that makes a difference. It is perfectly fine, one might contend, to contrast physical possibilities with physical impossibilities. But one cannot contrast logical possibilities with logical impossibilities. For a contrast class to be apt with respect to a particular why-question, we need to be able to make sense of the two situations we are contrasting. When one of the situations is a physical impossibility, we can still make perfectly good sense of the situation in question because it is constrained by the laws of logic. When one of the situations is a logical impossibility, we cannot make good sense of that situation; logical impossibilities are completely incoherent.

There are two ways to take this charge of incoherence. First, the charge might be that logical impossibilities are unimaginable and so we simply cannot conceptualise the contrast class at issue. Second, the charge might be that logical impossibilities are trivial, in this sense: everything and its negation is true in any situation featuring a logical impossibility.

We can respond to the first charge via the second. We need not suppose that logical impossibilities are trivial in the relevant sense. A logically impossible situation is trivial only if it features a contradiction and the situation itself is closed under the consequence relation of an explosive logic such as classical logic.[6] It would be rather odd, however, to suppose that when entertaining logically impossible situations, the notion of logical consequence at issue is that of an explosive logic. Classical logic is simply not suited to such applications.

Of course, we need to use some suitable consequence relation in our reasoning in order to see what follows from our inconsistent situation and what does not, but we don't need the situation itself to be closed under any consequence relation, and certainly not under a classical one. With respect to reasoning through an inconsistent situation, there are a number of non-explosive logics that fit the bill here. Some will be better than others for such purposes but all will be better than explosive logics such as classical logic. So with some non-explosive logic in hand we can specify the sense in which reasoning about an impossible situation is coherent. This, in turn, helps us to answer the charge of inconceivability. The very same imaginative capacities that we use to think through a logical possibility can be used to think through a logical impossibility. The only difference being, as discussed in the previous section, is that the relevant act of imagining must behave according to a different style of logical consequence.

[6] An explosive logic is one in which everything follows from a contradiction. A paraconsistent logic (or non-explosive logic), on the other hand, is one in which there is some proposition that does not follow from a contradiction.

There is a second, more flat-footed response to the charge that we cannot imagine impossible situations. The response is to point out that we do this all the time. We do this in logic, when we are reasoning through what would happen were various aspects of classical logic to be relaxed. We also do this in our day-to-day lives. Our imaginative capacities are liable to error and infelicity that infect the imaginings themselves. Often the things we take ourselves to be imagining are not strictly consistent; they are not crisp logical possibilities. They are messy, incomplete and inconsistent.

3.3 Ring-fencing

So far we have looked at two objections against the explanatory potential of logical impossibilities. But some may feel that there is a deeper problem that we haven't addressed. The worry is this: explanations by strict logical impossibility don't work like other explanations by impossibility. For example, consider an explanation by physical impossibility. Such an explanation works by showing us that there are possibilities that are inaccessible to us due to the actual laws of nature. The relevant possibilities are, to use a term from Smith, 'ring-fenced': there is a forbidden zone in modal space that the laws of nature prevent us from entering.

The same broad picture does not apply to the logical case. There are no possibilities in which a violation of a logical law happens. The logical laws do not 'ring-fence' any genuinely possible situations; there is no forbidden zone that the laws of logic prevent us from entering. This appears to be the basis of Smith's worry with explanations by logical impossibility. Speaking about the time-travel case again, Smith writes:

> This is a case of the type in which no (further) explanation of failure is required. There are no scenarios at all — no points in logical space — satisfying the description 'a time traveller [kills grandfather]'. There is no forbidden zone and hence no need or even possibility of an explanation of why the time traveller does not enter 'it'. Whatever happens, it won't be [killing grandfather] because no scenario at all satisfies that description. The reason for this is that the description is self-contradictory (e.g. it involves the time traveller permanently dying at 20 and also being alive at 40). So the crucial point here is that there is no forbidden zone. This is completely different from saying that there is one, but 'laws of logic' prevent us entering it. (Smith 2017, pp. 161–162)

By now, our response to this kind of worry should be obvious. We deny that there is nothing beyond the boundary of logical possibility. Beyond the boundary of logical possibility there are impossibilities. The logical laws do fence off a forbidden zone, the zone of situations that violate the logical laws in one way or another. Thus there are scenarios satisfying the description 'a time traveller kills their own grandfather before their father was conceived', they are just not possible scenarios. The laws of logic ring-fence the relevant scenarios, as depicted in Fig. 1.

One might respond that construed as explanations that rule something out, explanations by logical impossibility seem less substantive than other explanations

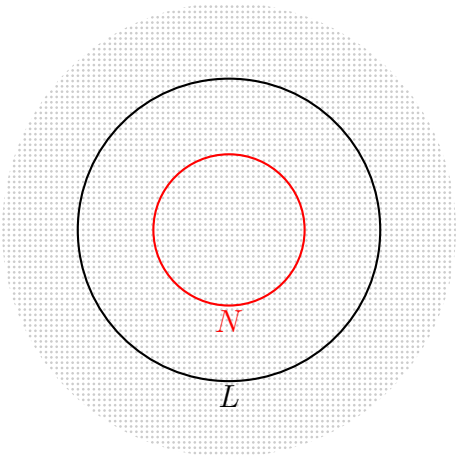

Fig. 1 Modal space again. N = the boundary of physical possibility, L = the boundary of logical possibility. Points outside of the boundary of L are impossibilities

by impossibility. The trouble comes this way. In the logical case, one might argue, what it means to say that some P is impossible is that P is ruled out by the laws of logic. Thus, to say that there is some impossible thing which is ruled out by the laws of logic is equivalent to saying that there is something that is ruled out by the laws of logic, which is ruled out by the laws of logic. Compare this to the physical case. In the physical case we can specify the thing being ruled out independently of the physical laws of nature. Thus, we can say that there is some possible P — where this just means that P is compatible with the laws of logic — and that P is ruled out by the physical laws of nature. Notice that in the physical case one is *not* making the trivial sounding claim that there is some P, which is ruled out by the laws of nature, that is ruled out by the laws of nature.

To see the problem more clearly, return again to the picture of modal space depicted in Fig. 1. We, of course, take there to be points beyond the boundary of the widest circle — these are our impossibilities. And so when we rule out some impossibility via the laws of logic, we are saying that the impossibility lies beyond the boundary of logical possibility and thus lies outside the largest circle. The problem is that the most natural account of what it is to be a logical impossibility is just this: to be a logical impossibility is to lie outside the largest circle. When we rule out some possibility via the laws of nature, we are saying that the possibility lies beyond the boundary of physical possibility, but within the circle of the logically possible. When we rule out a physical possibility, the ruling out seems substantive precisely because we have not defined whatever it is we are ruling out in terms of being beyond the circle corresponding to the physical laws. When we rule out a logical possibility, the ruling out *doesn't* seem substantive because we have defined what we are ruling out in terms of that thing being beyond the largest circle.

Now, in a certain sense, it is easy to provide a positive specification of the situation that is being ruled out. Take the time travel case again. We can simply say that the situation being ruled out is one in which Tim travels backwards in time and

manages to kill his grandfather before his mother was conceived. What we are looking for, however, is a positive specification of a different kind: of the modality of the situation. We are on the guard against triviality. We don't want to specify the modality of the situation by saying that it is logically impossible, and then go on to say that it is ruled out by the laws of logic, since that doesn't seem informative. What we want, rather, is a way to talk about the modal status of the situation so that the further claim that the situation is ruled out is meaningful.

We are assuming that Smith's notion of ring-fencing requires something along these lines. When we say that some situation is ruled out we draw a new line in the space of situations, a line that is supposed to give us new information. For the drawing of this new line to be informative, it must provide information that goes beyond any that we already have available to us in setting up the explanatory question in the first place. If, however, setting up the explanatory question requires already identifying some situations as logically impossible ones, going on to say that those situations are ruled out fails to yield new information. What we need, then, is a way to set up the space of situations against which the explanatory question can be asked and answered that does not require already saying that certain situations are logically impossible ones.

Again, this is analogous to the physical case. We can say that there are logically possible situations that are ruled out by the physical laws. When we do this, our initial identification of the modality of the situation is not simply a recapitulation of the claim that the situation is physically impossible. Rather, the claim that it is logically possible gives us information that is distinct to the information that is given by the relevant act of ruling out. The worry in the case of logical impossibilities is that there is no way to make the ruling out of a logically impossible situation give information that goes beyond any initial modal specification of the situation needed to set up the explanatory question being asked.

To address the problem and bring the logical and physical cases back into parity, we must provide a way to specify what's being ruled out in the logical case independently of the laws of logic. But this is straightforward to do using the second presupposition offered in Sect. 2: the assumption that impossibilities are structured. As noted, we are assuming that the impossibilities are closed under the non-classical logic R. With R in hand we can modify the picture of modal space in Fig. 1. We keep everything the same with respect to the physical laws of nature, and we retain a ring for logical possibility which has, nested within it, the ring for physical possibility. We now refine the picture by adding a new outer ring. This outer ring corresponds to the laws of R. In the space between the ring corresponding to the laws of classical logic and the laws of R we find situations that are compatible with R, but that are ruled out by the classical laws. Indeed, any situation involving a contradiction, but that does not violate the laws of R, will lie in this space. Situations violating one of R's axioms, such as the conjunctive axiom $A, B \vdash A \wedge B$ or the modus ponens axiom $A, A \rightarrow B \vdash B$ will lie beyond the outer-ring of R, since these will violate the laws of classical logic as well (see Fig. 2).

Are the situations that lie in the space between the laws of classical logic and the laws of R possibilities or impossibilities? The answer is that they are both: the situations are logical impossibilities, in so far as they are ruled out by the laws of

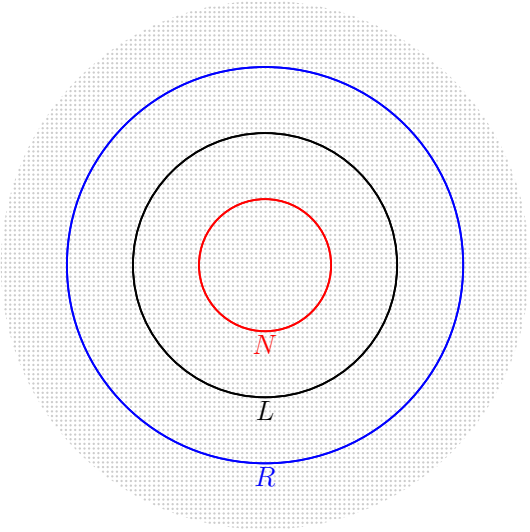

Fig. 2 Hyper Modal Space. N = the boundary of physical possibility, L = the boundary of logical possibility, R = the boundary of R-possibility, which are logical impossibilities

classical logic, and we take the laws of classical logic to govern the actual world. The situations are R-possibilities, however, in that they are compatible with the laws of R. Thus, if we lived in a world in which the laws of logic were R laws, then we would say that the R-possibilities are logical possibilities. Given that we don't live in such a world, the R-possibilities are not logical possibilities. They are logical impossibilities.

We can now specify the impossibilities that are ruled out by the classical laws of logic independently of the laws that are doing the ruling out. Indeed, once we have modified the broad approach to modal space in the above manner, the picture is exactly analogous to the situation with the laws of nature. Recall that in the physical case, the relevant notion of ruling out can be understood as follows. There is some P, which is possible — and thus compatible with the laws of logic — which is ruled out by the laws of nature. When P is ruled out by the laws of logic, we can now understand this as follows: there is some P which is R-possible — i.e., compatible with the laws of R — but which is ruled out by the laws of logic. Thus, we are no-longer forced to say the trivial sounding "there is a P, which is ruled out by the laws of logic, that is ruled out by the laws of logic". Of course, we *can* still say this if we want to. But we can also say this in the physical case. The point, however, is that in addition to this trivial way of specifying the ruling out procedure, there is a non-trivial way to specify it in both cases. Moreover, the structure of the two cases is now the same, so there is little reason to suppose that one type of 'ruling out' is explanatorily substantive, while the other type is not.

Of course, in the resulting picture, we are forced to say that the laws of R hold actually, just as we are forced to say in the physical case that the laws of logic hold

actually. This is a direct corollary of the fact that the actual world is contained within both the ring corresponding to the laws of logic — which, we will assume for the sake of argument, are the classical laws — and the ring corresponding to the laws of *R*. But, in fact, it *is* the case that the laws of *R* hold actually, it is just not the case that the laws of *R* are the *logical* laws. This is analogous to the way in which the laws of logic hold actually but are not the physical laws of the actual world. They correspond to another, weaker constraint.

The relationship between the physical laws and the laws of logic is exactly the same. There is no problem with the idea that there are situations that are compatible with the logical laws but are not compatible with the laws of nature. Similarly, there is no problem with the idea that there are situations that are compatible with the laws of *R* but that are not compatible with the laws of logic. It is these situations that the laws of logic rule out, just as the laws of nature rule out the classical, but physically impossible, situations.[7]

4 Impossibilities again

This concludes our defense of the idea that explanations by logical impossibility can be genuine explanations. As discussed, the defense presupposes that there are logical impossibilities in at least some sense. Does the appeal to impossibilities itself pose a problem for our argument? One might think so. If we are now countenancing impossibilities, it might be thought that we need an explanation of why we find ourselves in a possible situation rather than an impossible one.[8] According to this line of argument, once impossibilities are on the table and we are without an account of why impossibilities can't be actual, our account of logical explanation fails. Consider, once more, the time travel case with which we began. We want to say that Tim can't kill his grandfather because it would be impossible for him to do so. But if there are impossibilities, then there are situations in which Tim succeeds in killing his grandfather. For the explanation to succeed, we need to say something about what prevents us from being in a situation where Tim can succeed in killing his grandfather.

We can differentiate between two different versions of this objection. The first version goes like this. Suppose that both possibilities and impossibilities exist, and are concrete. Then for all we know, the world we live in is a logically impossible one. Indeed, for all we know the world might be only very minimally impossible, in the sense that in some far-flung corner of the universe there is a contradiction lurking, but everything around here is perfectly well-behaved (both physically and

[7] The analogy might run even deeper, if one thinks that theory choice in logic is broadly an empirical matter. The logically possible is just that which is specified as such by the preferred logical theory; the physically possible is just that which is specified as such by the preferred physical theory. In both cases, the decision about which is the preferred theory is made in part on empirical grounds. See Putnam (1979) for a view along these lines.

[8] See Smith (to appear, 2017) for the source of this argument.

logically). The trouble is that all of the evidence we have gathered to date cannot rule out such a hypothesis about the world.

This first version of the objection, however, is not a concern for our position. We are not committed to the claim that impossibilities exist and that they are concrete. The objection only really gets going, however, if a concrete realist approach to impossibilities is presumed. For if impossibilities are not concrete but actuality is, then we know that we are not in an impossible situation. Similarly, if impossibilities do not exist, but actuality does, then we know for Cartesian reasons that we are not in an impossible situation (i.e., because we know that we exist).

The second version of the worry does not hang on any particular metaphysics. The worry is best framed by thinking about the physical laws of nature. We have evidence that the physical laws are thus and so, gathered from the progress of science. When we appeal to the physical impossibility of some fact in an explanation, we have good reason to take the appeal seriously, because we have substantial evidence in favour of the impossibility in question. When it comes to logical impossibilities, however, one might worry that we don't have the same evidential basis for these impossibilities. We don't have anything like scientific or empirical evidence that, say, the law of non-contradiction is true.

There are two responses available to this version of the worry. First, we do have evidence of a kind that the 'laws of logic' are a particular way. No one has ever observed a contradiction, and that's evidence of a kind that there aren't any. Of course, one might dispute the evidence. Priest, for instance, argues that paradoxes such as the liar paradox and the sorites paradox are instance of genuine inconsistency in the actual world (Priest 2002). More routinely, there are simple cases of conflicting evidence and conflicting scientific theories. These too might give us reason to suppose that the actual world is inconsistent (Colyvan 2002). Even if one thinks that there is evidence against the law of non-contradiction, however, one still thinks that what the logical truths are is a fact that is sensitive to evidence. That's all we really need for our purposes, since then we can discover the laws of logic via broadly epistemic methods, similar to the way we discover the physical laws.[9]

If one does not think that the laws of logic are established by evidence, then presumably that is because one thinks they are *a priori*. Assuming we have access to this *a priori* knowledge, then we have what we need for cases of explanation by logical impossibility. It doesn't matter, for our purposes, what the source of knowledge regarding the logical truths ultimately is. Our point is that so long as we have such knowledge, then it can be used to scaffold a certain kind of explanation. To be clear, we don't mean to underplay the question of how we know what we know about logic. This is a deep question. But it is also largely orthogonal to our project.

No doubt there is more to say about these issues but for now we see our account of logical explanation sitting very comfortably with similar accounts appealing to, for example, physical impossibility.

[9] See Bueno and Colyvan (2004) for an account along these lines.

References

Arntzenius, Frank, & Maudlin, Tim. (2002). Time travel and modern physics. In Craig Callander (Ed.), *Time, reality and experience* (pp. 169–200). Cambridge, MA: Cambridge University Press.

Baron, Sam, & Colyvan, Mark. (2016). Time enough for explanation. *Journal of Philosophy*, *113*, 61–88.

Baron, Sam, & Colyvan, Mark. (2019). The End of Mystery. *American Philosophical Quarterly*, *56*(3), 247–264.

Baron, Sam, Colyvan, Mark, & Ripley, David. (2017). How mathematics can make a difference. *Philosophers' Imprint*, *17*, 1–19.

Bernstein, Sara. (2016). Omission impossible. *Philosophical Studies*, *173*, 2575–2589.

Berto, Francesco, French, Rohan, Priest, Graham, & Ripley, David. (2018). Williamson on counterpossibles. *Journal of Philosophical Logic*, *47*, 693–713.

Bjerring, J. C. (2014). On counterpossibles. *Philosophical Studies*, *168*, 327–353.

Brogaard, Berit, & Salerno, Joe. (2013). Remarks on counterpossibles. *Synthese*, *190*, 639–660.

Bueno, Otávio, & Colyvan, Mark. (2004). Logical non-apriorism and the law of non-contradiction. In G. Priest, J. C. Beall, & B. Armour-Garb (Eds.), *The law of non-contradiction: New philosophical essays* (pp. 156–175). Oxford: Oxford University Press.

Carroll, John W. (2010). Context, conditionals, fatalism, time travel and freedom. In Joseph Keim Campbell, Michael O'Rourke, & Harry S. Silverstein (Eds.), *Time and identity* (pp. 79–93). Cambridge, MA: MIT Press.

Colyvan, Mark. (2002). The ontological commitments of inconsistent theories. *Philosophical Studies*, *141*, 115–123.

Dowe, Phil. (2007). Constraints on data in worlds with closed timelike curves. *Philosophy of Science*, *74*, 724–755.

Elga, Adam. (2001). Statistical mechanics and asymmetry of counterfactual dependence. *Philosophy of Science*, *68*, S313–S324.

Fine, Kit. (1975). Review of "counterfactuals". *Mind*, *84*, 451–458.

Gorovitz, Samuel. (1964). Leaving the past alone. *Philosophical Review*, *73*, 360–371.

Hájek, A. Most counterfactuals are false. Unpublished manuscript.

Hoeltje, Miguel, Schnieder, Benjamin, & Steinberg, Alex. (2013). Explanation by induction? *Synthese*, *190*, 509–524.

Horwich, Paul. (1987). *Asymmetries in time*. Cambridge, MA: MIT Press.

Ismael, Jennan. (2003). Closed causal loops and the bilking argument. *Synthese*, *136*, 305–320.

Jago, Mark. (2015). Impossible worlds. *Noûs*, *49*, 713–728.

Kim, Jaegwon. (1973). Causes and counterfactuals. *Journal of Philosophy*, *70*, 570–572.

Kment, Boris. (2014). *Modality and explanatory reasoning*. Oxford: Oxford University Press.

Lewis, David. (1973a). Causation. *Journal of Philosophy*, *70*, 556–567.

Lewis, David. (1973b). *Counterfactuals*. Oxford: Blackwell.

Lewis, David. (1976). The paradoxes of time travel. *American Philosophical Quarterly*, *13*, 145–152.

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

Lewis, D. (ed.) (1986). Causal explanation. In *Philosophical papers II* (pp. 214–240). Oxford: Oxford University Press.

Lewis, David. (1987). Events. *Philosophical papers II* (pp. 242–270). Oxford: Oxford University Press.

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

Moss, Sarah. (2012). On the pragmatics of counterfactuals. *Noüs*, *46*, 561–586.

Nolan, Daniel. (1997). Impossible worlds: A modest approach. *Notre Dame Journal of Formal Logic*, *38*, 535–572.

Nolan, Daniel. (2014). Hyperintensional metaphysics. *Philosophical Studies*, *171*, 149–160.

Priest, Graham. (2002). *Beyond the limits of thought*. Oxford: Oxford University Press.

Putnam, Hilary. (1979). The logic of quantum mechanics. In *Mathematics, matter and method: Philosophical papers* (2nd ed., Vol. 1, pp. 174–197). Cambridge: Cambridge University Press.

Restall, Greg. (1997). Ways things can't be. *Notre Dame Journal of Formal Logic*, *38*, 583–596.

Riggs, Peter J. (1997). The principal paradox of time travel. *Ratio*, *10*, 48–64.

Ripley, David. (2012). Structures and circumstances. *Synthese*, *189*, 97–118.

Schnieder, Benjamin. (2008). On what we can ensure. *Synthese*, *162*, 101–115.

Schnieder, Benjamin. (2011). A logic for 'because'. *The Review of Symbolic Logic*, *4*, 445–465.

Schnieder, Benjamin. (2016). In defence of a logic for 'because'. *Journal of Applied Non-Classical Logics*, *26*, 160–171.
Smith, Nicholas J. J. (1997). Bananas enough for time travel? *British Journal for the Philosophy of Science*, *48*, 363–389.
Smith, Nicholas J. J. (2017). I'd do anything to change the past (but I can't do 'that'). *American Philosophical Quarterly*, *54*, 153–168.
Smith, N. J. J. (to appear). Against impossible worlds.
Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Oxford: Blackwell.
Tsohatzidis, S. (2015). A problem for a logic of 'because'. *Journal of Applied Non-Classical Logics*, *25*, 46–49.
Vander Laan, David. (1997). The ontology of impossible worlds. *Notre Dame Journal of Formal Logic*, *38*, 597–620.
Williamson, Timothy. (2013). *Modal logic as metaphysics*. Oxford: Oxford University Press.
Woodward, James. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
Woodward, James, & Hitchock, Christopher. (2003). Explanatory generalizations, part 1: A counterfactual account. *Noûs*, *37*, 1–24.