Against the Standard Solution to the Grandfather Paradox
[Penultimate Draft: Forthcoming at *Synthese*]

Abstract

1000 time-travelers travel back in time, each with the intention of killing their own infant-self. If there is no branching time, then on pain of bringing about a logical contradiction, all must fail. But this seems inexplicable: what is to ensure that the time-travelers are stopped?

For a time, this *inexplicability objection* was thought to provide evidence that there is something incoherent about the possibility of backwards time travel in a universe without branching time. There is now near-consensus, however, that the objection has no bite: there is nothing inexplicable about the mass failure. Lewis, Sider and Ismael independently argue that since it is built into the description of the class of cases considered that the time-travelers must fail – and so we consider only unsuccessful attempts – there is no mystery. Smith argues that the absence of possible worlds at which auto-infanticide is committed suffices as a complete explanation for the failures. And Baron and Colyvan maintain that available causal and logical explanations jointly account for everything that needs accounting for.

I argue that these are wrong. There is remaining, problematic inexplicability. For backwards time travel not to lead to logical contradiction, something would need to do logic's bidding, after all.

Keywords: Time Travel, Grandfather Paradox, Closed Causal Loops, Backwards Causation, David Lewis, Explanation

1. The Puzzle

Is backwards time travel metaphysically possible in a universe without branching time? Since David Lewis's seminal (1976), it has been widely accepted that the answer to this question is *yes*: there is nothing incoherent about backwards time travel, even if time is one-dimensional. Here I argue that the Lewis-style reply is wrong, and that the problem brought out by the grandfather paradox is more serious than generally thought.

The *Grandfather Paradox* refers to the following classic puzzle. If backwards time travel were possible, a time traveler could travel back to when her grandfather was very young and kill him well before his child, the time traveler's parent, was conceived. Surely this would be possible: there would be nothing, in principle, to stop her. On the other hand, it is clear that it could not be possible. For killing her grandfather before he helped conceive her parent would give rise to a contradiction,

indeed a host of contradictions: the time traveler both is and is not born, her grandfather both did and did not live past childhood, the time traveler both did and did not kill her grandfather when he was young (for if she did, then she was never born, in which case she didn't), etc. So it seems that we have two choices: deny that it would be possible for the time traveler to kill her grandfather were she to travel back in time, or deny that backwards time travel is possible. But since it seems clear that a time traveler could kill her own grandfather, we must deny the possibility of backwards time travel.

There are different variations of this objection. Here I focus on a variation that is sometimes referred to as the *inexplicability objection*. To explain the objection, we will need a time traveler, call her $Timna_{OLDER}$ ("$Timna_O$", for short). Instead of having $Timna_O$ try to kill her grandfather, she will travel back in time to try to kill her own infant self, $Timna_{YOUNGER}$ ("$Timna_Y$"). When $Timna_O$ steps out of her time machine in the past, she finds her helpless younger self alone in a crib. $Timna_O$ is highly trained in both Karate and Taekwondo and has a sword, dagger, and fully loaded gun as backup. $Timna_O$ attempts to kill the child. What happens? If we suppose for the sake of argument that backwards time travel *is* possible, we will have to accept that although it seems that $Timna_O$ should have no problem killing $Timna_Y$, in fact she cannot do so: something will have to go wrong. $Timna_Y$ will have to survive. Maybe $Timna_O$ becomes squeamish about the idea of creating a logical contradiction at the last moment and cannot bring herself to go through with the plan. Maybe she slips on a banana peel. Maybe she accidentally kills the wrong child. Whatever happens, $Timna_O$ must fail. But this is inexplicable. For there is nothing in place to ensure that she fails. As David Lewis puts it, "The forces of logic will not stay [her] hand! No powerful chaperone stands by to defend the past from interference." (1976:149). Of course, coincidences happen. It is certainly *possible* for $Timna_O$ to happen to fail for some fluky reason or other. But coincidences are just that: coincidental. What seems inexplicable is that the pertinent causal chain *must* terminate in her failure.

The problem becomes more striking still when we consider a scenario in which not merely one time traveler, but many – perhaps 1,000 – all travel back in time, each with the intention of committing auto-infanticide. Although the conditions are ripe for the time travelers to succeed (they are all capable assassins confronting helpless infants), something must go wrong in every single case. Not one of the time travelers can succeed in the task. How can this be? There are, it seems, two good options for responding to this puzzle. We can try to explain away the (appearance

of) inexplicability; or else we can accept that, at the very least, the paradox provides us with compelling *evidence* that there is something incoherent about backwards time travel in a universe without branching time (albeit not conclusive proof of its impossibility).

2. The Standard Solution

Lewis (1976) was the first to defend time travel against the inexplicability objection. There have since been a number of articles responding to it approximately as he does. The response, which is widely accepted as correct, is to take the first option above and deny that there is actually anything inexplicable or extraordinary about the fact that something or other will cause each time traveler who tries to commit auto-infanticide to fail. And if there is nothing extraordinary about it, then the fact that each must fail gives us no reason whatsoever to reject the possibility of backwards time travel.

But how can there be nothing extraordinary about the fact that every single time traveler who tries to commit auto infanticide will fail? How can the appearance of inexplicability be explained away? Lewis's solution is to invoke the fact that 'can'-claims are context sensitive. In a context in which we only consider facts about the time before Timna_O's attempt on Timna_Y's life – facts, for example, about Timna_O's abilities and opportunities at the time – it will turn out that Timna_O *can* kill Timna_Y, where by this it is meant that Timna_O's killing of Timna_Y is compossible with the relevant pre-attempt facts. This is a standard way to evaluate a 'can'-claim, and many of the 1000 backwards time travelers can kill their younger selves in this sense. It is this sense of 'can' that we use when we use reasoning like, 'of course Timna_O can kill Timna_Y. The latter is just a helpless infant and there's nothing to stop Timna_O…' We shift to a different use of 'can', however, when we think about the killing being logically impossible. In particular, we hold fixed facts like that Timna_Y survived long enough to become an adult and travel back in time. Once we recognize that we are holding these additional facts fixed, the inexplicability vanishes: the fact that Timnao cannot kill Timna_Y *given that* Timna_Y survived the attempt is no more inexplicable than is the fact that I cannot eat a sandwich for lunch tomorrow *given that* I will eat (only) salad.

Along similar lines, Jenann Ismael (2003) makes the case for explicability by comparing the time traveler scenario, which I dub '*Auto-Infanticide*', to a scenario in which Ismael tries and fails to reach her mother by telephone (call this '*Telephone*'). There is nothing odd or mysterious about the

fact that every time Ismael attempts to reach her mother by telephone *but fails*, there is some causal explanation for the failure. It may be unlikely or coincidental for Ismael's phone call to fail to go through; nonetheless, it is expected that *each time it does fail*, there is an explanation. By attending to just those cases (however rare) in which the call does not go through, we selectively attend to just those cases in which something or other has caused the phone call attempt to fail.

Similarly, Ismael argues, when we consider cases in which there is a time traveler from the future – alive and in position to attempt auto-infanticide – we are already considering only cases in which that time traveler fails. For "…when we describe a self-defeating causal chain, we sneak in, under the guise of the first event, a description of the *last* which is incompatible with the success of the operation as a whole…We consider only unsuccessful cases." (2003: 308, her emphasis). Timna being alive as an adult entails that she survived past her childhood. She has survived any attempts made on her life when she was a child, including any made by her older self. And just as it makes no difference how many times Ismael tries *and fails* to call her mother – each time there will be some causal explanation of her failure – similarly, each time a time traveler tries (and indeed, fails) to commit auto-infanticide, there will be some causal explanation for this failure as well. Finally, just as there is nothing spooky or noteworthy about this in the case of Ismael's failed attempts at mom-calling, likewise there is nothing spooky or noteworthy about it in the case of the auto-infanticidal time travelers, regardless of how weird or coincidental the failure appears each time.[1]

3. A Disanalogy?

While it appears, and indeed is widely accepted, that the puzzle has been solved, I now argue that it has not been. Let us suppose that Timna_O's attempt to kill Timna_Y occurs at some particular time, t1. I will use "x $\text{can}_{[<t1]}$ *Φ*" to mean that x *Φ*-ing is compossible with the relevant *pre-t1* facts, and "x $\text{can}_{[\infty]}$ *Φ*" to mean that that x *Φ*-ing is compossible with all the relevant facts, including facts at and after t1. We have, then, the following two claims:

(i) Timna_O $\text{can}_{[<t1]}$ kill Timna_Y.

(ii) Timna_O $\text{can}_{[\infty]}$ kill Timna_Y.

[1] Ted Sider (2002) argues similarly.

I agree with Lewis that the inexplicability vanishes if (i) is true and (ii) is false. And while I agree that (ii) is false, I disagree that (i) is true. Timna_O's successful killing of Timna_Y is *not* compossible with the relevant pre-t1 facts. She *must* be caused to fail, even when we don't hold fixed facts at or after t1. If that is right, then that something will be in place to cause her to fail (a banana peel, a jammed gun, etc.) remains inexplicable – or so I will argue. I have two premises, then, to defend. First, that Timna_O $\text{cannot}_{[<t1]}$ kill Timna_Y. Second, that if that is right – i.e., if (i) is false – then *Auto-Infanticide* is problematically inexplicable. Here is the plan. In the remainder of this section and in §4, I will argue that the effectiveness of the analogy between *Auto-Infanticide* and cases like *Telephone* depends crucially on (i) being true. That is, *Telephone* is only explicable because it is the case that, given that the call is placed at t1, Ismael $\text{can}_{[<t1]}$ reach her mother. In §5 I argue that Timna_O $\textit{cannot}_{[<t1]}$ kill Timna_Y, and that we therefore cannot account for the explicability of *Auto-Infanticide* in the way that Lewis, Ismael, and Sider do. In §6 and §7 I consider different ways theorists have tried to defuse the inexplicability worry, and I argue that they are not successful, either.

To better understand the explicability of *Telephone*, it will be helpful to consider an additional scenario which is explicable in the way that *Telephone* is but which, like *Auto-Infanticide*, involves backwards time travel.

Bicycle: Ernie_O never had a bicycle as a child. He travels back in time, meets his younger self, Ernie_Y, and decides to buy him a bicycle. Since Ernie_O did not have a bicycle as a child, something or other must cause Ernie_O to fail in his task. Indeed, all time travelers who did not have bicycles as children will fail if they attempt to travel back in time and give their childhood selves bicycles.

Why is it explicable that something – perhaps something that appears fluky and coincidental at the time – will happen to cause the call to not go through each time Ismael tries and fails to reach her mother? And why is it explicable that every time traveler who did not have a bicycle as a child and who travels back in time and attempts to give his or her younger self a bicycle will fail?

There is a feature, shared by *Telephone* and *Bicycle,* which, I maintain, accounts for the scenarios' explicability. Here is a first attempt at it.

Contingent Failure: Although the attempt in question failed, it was the case from a temporal perspective prior to the failure that it could have succeeded.

When I say of a non-occurring event that it *can* occur, or *could have* occurred, or that its nonoccurrence is contingent from the perspective of some time, t, I mean that the event is metaphysically compossible with all (or nearly all[2]) pre-t facts. I specify the temporal perspective from which the event can occur to distinguish it from the trivial way in which it cannot occur from the perspective of a time *after* it does not occur. If we hold fixed the fact that Ismael's call failed, for instance, then it is necessary that it did. That is, at all the worlds at which Ismael tries to reach her mother *and the call does not go through*, Ismael's attempt fails. Similarly, Ernie$_O$ fails to give Ernie$_Y$ a bicycle at all worlds at which Ernie$_Y$ never had a bicycle. But, supposing these failures both occur at t1, and holding only pre-t1 facts fixed, neither failure was metaphysically necessary. There are nearby possible worlds at which Ismael's phone call goes through. Likewise, from a temporal perspective prior to Ernie$_O$'s failed attempt to give Ernie$_Y$ a bicycle, it was not metaphysically required that he fail. There are nearby possible worlds at which Ernie$_O$ tries to give Ernie$_Y$ a bicycle and succeeds. Of course, at these worlds Ernie$_O$ *does* have a bicycle as a child.

The importance of the contingency of the failure given an appropriate temporal perspective is especially clear in examples involving backwards causation. Consider *Bicycle* again. Ernie did not have a bicycle as a child. At first sight it might seem inexplicable that, since he did not have a bicycle, if Ernie$_O$ were to travel back in time and try to give Ernie$_Y$ a bicycle, he would have to be caused to fail in some way. What is to ensure that he does? *It is because of the contingency of his failure that we don't need to answer this question*: prior to t1, it is not the case that he has to fail. Nothing needs to ensure that it happens. It is just that he contingently (from a pre-t1 perspective) *does* fail. It is, in part, because Ernie$_O$ happens to fail at t1 that Ernie never had a bicycle as a child. It would be inexplicable if Ernio$_O$ had to fail if we didn't hold fixed anything at or after t1.

I will argue in the next section that *Auto-Infanticide* does not share this feature with the other cases. From *no* perspective on the timeline is it metaphysically possible for Timna$_O$ to kill Timna$_Y$. And it was this feature that made the respective failures in *Telephone* and *Bicycle* explicable. If it is

[2] If x does *not* Φ at w, then x's Φ*-ing* may *not* be compossible with the conjunction of w's laws and the set of *all* pre-t facts at w; especially if w's laws are deterministic. In that case, we can take $can_{[<t]}$ to mean that x Φ*-ing* is compossible with as many pre-t w-facts as possible, allowing only the minimal change/s required for x to Φ at w.

not the case that Timna$_O$ merely happens to fail due to causal happenstance, then the question, 'what is there to ensure that Timna$_O$ fails?', still requires answering, and the inexplicability remains.

4. *Contingent Failure* Improved

In both *Telephone* and *Bicycle*, we considered whether the agent could have succeeded in his or her respective task from a temporal perspective prior to t1, the time of the attempt. Since in both cases they could have, the actual failures could be attributed to ordinary contingent causal accident and so were deemed explicable. As we will see in a moment with the example *Light Switch*, however, there are some scenarios involving backward causation in which a particular event *must* be caused to occur from a temporal perspective prior to its own occurrence, but where this is still explicable. Examples of this kind show that we'll need to be more precise about the temporal perspective from which the occurrence of an event should be contingent if it is to be explicable.

Suppose there is a switch that, if flipped at t2, makes a light turn on at an earlier time, t0. For simplicity we can suppose that the only way for the light to turn on at t0 is for the switch to be flipped at t2. Call this scenario *Light Switch*. One who has all the relevant facts can epistemically infer from the light coming on at t0 that the switch will be flipped at t2. Indeed, if the light comes on at t0 – still assuming that its cause is the switch being flipped at t2 – then we might be inclined to think that at t0 it is already the case that the switch will (have to) be flipped at t2. Given that the light has already come on at t0 and that its cause is the switch being flipped at t2, were someone to try at t1 to make it such that the switch is not flipped, that person would fail. It seems that they would have to fail even before they tried. And yet, I take the (real) lesson of the Lewis-Ismael-Sider line to be, in part, that there need be nothing inexplicable about this failure: the failure *happens* to occur at t1, and it is only because it does that the light comes on at t0. Were it *not* to (happen to) occur, the light would not have come on. The reason this failure is explicable is that there is still some important sense in which the failure was contingent, even though it already had to occur prior to the time of its occurrence at t1.

So in what sense did the failure merely happen to occur? I suggest it is this: the switch flipping is contingent *if the causal effects of the switch flipping are not held fixed*. Alternatively, we can express the same idea as follows: the switch flipping is contingent from a temporal perspective that is prior to both the switch flipping *and* to all of its causal consequences. Let us suppose that the light turning

on at t0 is the earliest effect of the switch being flipped. In that case, it is possible for the switch to not be flipped (and thus, for the light to not turn on) from a temporal perspective that is prior to t0. There are nearby possible worlds at which the switch is not flipped at t2 and, because of this, the light does not turn on at t0. Because the complete causal chain from the switch flipping to the light turning on is contingent, nothing has to be in place to ensure that the switch is flipped at t2. Either something happens to make it such that the switch is flipped, in which case the light turns on at t0; or not, in which case the light stays off. Our discussion of *Light Switch* indicates that *Contingent Failure* should be modified in the following way:

> **Contingent Failure'**: Although the attempt in question failed, it was the case from a temporal perspective prior to the failure *and to its earliest causal consequence* that it could have succeeded.

Return to *Auto-Infanticide*. If Ismael is right that *Auto-Infanticide* is explicable in the way that the other cases are, then it should be possible for Timna_O to succeed at killing Timna_Y from a temporal perspective that is prior to not only Timna_O's failure to kill Timna_Y, but also to the earliest causal effects of the failure. Timna_O's arrival by time travel at t0 is causally downstream from Timna_O's failure to kill Timna_Y since, were she to not fail, Timna_O would not be alive nor able to time travel to t0. And if that is right, then for *Auto-Infanticide* to be explicable in the way that *Light Switch* and the others are, it must be the case that Timna_O can succeed at killing Timna_Y from a temporal perspective that is prior to t0. If she $\text{can}_{[<t0]}$ succeed, then her failure can be attributed to ordinary causal happenstance: she *happened* to fail, and so Timna_O was alive and able to time travel to t0.

The question, then, is this: $\text{can}_{[<t0]}$ Timna_O kill Timna_Y? In other words, is it the case that if Timna_O fails, it is because she merely happens to? Is the failure contingent from a temporal perspective prior to t0? And does Timna_O merely happening to fail thus (partially) account for how Timna_O is able to time travel to t0 in the first place? As in *Light Switch*, we can test the contingency of Timna_O's failure by testing the contingency of the causal chain as a whole. We do this by investigating if it is possible to "pick up" the whole chain from the cause – Timna_O's failure to kill Timna_Y — to the effect – Timna_O arriving via time machine at t0. If it is the case that Timna_O

contingently happens to fail, it should be the case that Timna_O $\text{could}_{[<t0]}$ have succeeded at killing Timna_Y if we allow for the effect of her failure, her time travel to t0, to vary when we vary the cause.

When we try to test for contingency we immediately run into a problem, however. If Timna_O hadn't time traveled to t0 she wouldn't have made the attempt to kill Timna_Y in the first place, let alone succeeded or failed. What then is indicated by Timna_O's presence at t0? Does it epistemically indicate that she contingently happened to fail, the way the light being on epistemically indicates that the switch happened to be flipped? Does it indicate that although she $\text{could}_{[<t0]}$ have succeeded, she didn't, and that's (partly) why she is at t0 – the way that, e.g., Ernie never having a bicycle as a child indicates that if Ernie_O tried to buy his younger self a bicycle he happened to fail (and that failure is partly why he never had a bicycle)? It does not. Timna_O's presence at t0 does *not* indicate that she contingently happened to fail if she tried to kill Timna_Y – i.e., that she *happened* to slip on a banana peel, or the gun *happened* to jam, etc. – since success would require her to first be at t0 as well. Of course, we know that she must fail. It is logically required that she does. The point is that her presence at t0 does not indicate that she contingently happens to fail. For the failure to be contingent, it must have been possible for it to *not* have occurred given that most pre-t0 facts are held fixed while the failure and its causal consequences are not.[3] It must be the case that the agent $\text{can}_{[<t0]}$ succeed.

5. $\text{Can}_{[<t0]}$ Timna_O kill Timna_Y?

So $\text{can}_{[<t0]}$ Timna_O kill Timna_Y? We've already seen that Timna_O's presence at t0 does not indicate that she happened to fail to kill Timna_Y if she tried, and so the case is already importantly different from the others. But from this does it follow that Timna_O $\text{cannot}_{[<t0]}$ kill Timna_Y? More to the point, does it follow that she didn't *contingently happen* to fail from a pre-t0 perspective? A bit more must be said to answer this question. In particular, its answer depends on whether there are already enough facts, pre-t0, to ground Timna_Y's identity to Timna_O. If there are enough such facts, then we need not hold fixed post-t0 facts for it to be the case that Timna_O cannot kill Timna_Y, and

[3] To test for contingency, we hold fixed as many pre-t0 facts as possible because we want to know if the respective agent can succeed at the relevant task (buying Ernie a bicycle, killing Timna_Y, etc.) at other worlds that are just about the same as *w* prior to t0.

the answer to the question in the heading of this section is "no". I thus turn my attention to the following question:

Q: Is there some pre-t0 fact or set of facts which is sufficient to ground Timna$_O$'s identity to Timna$_Y$?

The kind of identity at issue here is the kind that would make it such that if Timna$_Y$ is identical to Timna$_O$, then the former's death is inconsistent with the latter's existence. We can call this *origination-identity*, since Timna$_O$ must originate or emerge out of Timna$_Y$ for Timna$_Y$'s death to be inconsistent with Timna$_O$ being alive. Which facts ground origination-identity is a difficult question. We might appeal to Kripke's necessity of origin thesis: if Timna$_Y$ and Timna$_O$ come from the same zygote, then they do so necessarily and the two are origination-identical at all possible worlds at which both exist. In that case it makes no difference what happens to Timna at or after t0: every world with both Timna$_Y$ and Timna$_O$ is a world at which one emerges from the other.[4] Since they originate from the same zygote by necessity, it is a necessary fact that the two share the same origin. And in that case, the answer to Q seems to be *yes*: Timna$_Y$ and Timna$_O$ are origination-identical, even without holding fixed any facts at or after t0. Because the same individual is rigidly designated by "Timna$_Y$" and "Timna$_O$" – just at different life stages – it is the case from *all* temporal perspectives that the older must fail to kill the younger if she time travels and tries. And if that is right, then assuming that Timna$_O$ is older, she does not merely happen to fail at t1. She had to fail, even without holding any future facts fixed. (As we've seen, it is *not* open to those endorsing the Lewis/Ismael line to reply that had Timnao succeeded then Timna$_O$ wouldn't exist in the first place, and thus that her presence at t0 just proves that she did happen to fail. This is not a good reply since, first, Timna$_O$ would also have to first be at t0 *in order to succeed* and make it such that Timna$_Y$ doesn't survive, and anyway, succeeding is itself a contradiction. One who does not exist does not succeed at anything.)

But what if the answer to Q is *no*? Suppose we adopt a view of identity according to which one's identity, even her origination-identity, is grounded in nothing less than that individual's entire temporally-extended worm. On this view, whether Timna$_O$ is origination-identical to Timna$_Y$ at t0

[4] Could it be that at some worlds Timna$_Y$ is older than Timna$_O$ and originates from her? If so, then at these worlds Timna$_O$ can kill Timna$_Y$ but Timna$_Y$ cannot kill Timna$_O$. The scenario we are considering is one where the older time traveler attempts to kill her younger self.

depends, in part, on what happens at and after t1. If there are *not* enough pre-t0 facts about Timna_O and Timna_Y to fix whether they are identical, then the explicability can be accounted for as follows: nothing needs to stop Timna_O from successfully killing Timna_Y. If she happens to be successful, then this will make it such that she is origination-identical with a different (actual) younger person, and so there will be no logical contradiction.

Let us momentarily grant the controversial assumptions about identity needed to run this line. Still, we confront a serious difficulty. If it is the case that Timna_O $\text{can}_{[<t0]}$ kill Timna_Y, as is required to solve the inexplicability in the way Lewis and Ismael claim to, then it must be the case, prior to t0, that there is more than one (younger) person at *w* who is a candidate for being identical with Timna_O. If Timna_Y is the only person who is such a candidate, then this strategy for handling the apparent inexplicability will not help: it will still be the case, prior to t0, that Timna_O must fail if she tries to kill Timna_Y. The only way it is possible for her to succeed is if the killing would make it such that Timna_O is origination-identical to someone else. But this requires that there be someone else alive with whom it is possible for her to be origination-identical, depending on what happens at and after t1. And indeed, everyone who might backwards time travel would need to have multiple younger people with whom they could stand in the relevant relation of identity, if necessary. This may already be weird enough, but it is not only a problem of *weirdness*: things get worse if we revise the scenario. It is not sufficient for there to only be *one* such additional person with whom each person can be identical, for suppose that the time traveler's scheme involves aiming to kill the additional candidate, too. Indeed, let us devise a scenario in which Timna_O's plan is to travel back to t0 and destroy the whole world. Suppose she tries to do this. Suppose 1000 time travelers try to do this. It does not matter how advanced the destroy-the-world technology: something would have to ensure that at least one (suitable) person survive per each time traveler so that each time traveler can be origination-identical with someone. It is logically required that something stop the time travelers from succeeding at killing everyone, and so the inexplicability remains.

There is another possibility to consider. What if there are alternative metaphysically possible ways for Timna_O to be present at t0 – ways which do not make it impossible for her to kill Timna_Y. For example, Timna_O could spontaneously materialize at t0, even if Timna_Y dies at t1. Is a possibility like this relevant to our discussion? Do we avoid the difficulties if the (only) reason that Timna_O

can$_{[<t0]}$ kill Timna$_Y$ is that she can, in theory, also arrive at t0 in some way other than by time travel? The answer to these questions, I think, is no: possibilities of this kind are not to the point. We can begin by noting the intuitive implausibility of such possibilities making a difference here: could it really be that backwards time travel is only metaphysically possible in a universe with one-dimensional time if (and because?) spontaneous materialization (or some suitable alternative) is possible, too (since, on the presently-considered proposal, it's the possibility of an alternative way of getting to t0 that can get us out of the problem)? For these very different kinds of potential possibilities to be conceptually tied in this way would be extraordinary, on its face. [5]

Indeed, I now argue that the inexplicability is not gotten rid of by bringing in alternative ways for Timna$_O$ to get to t0. We can, in fact, set such possibilities aside. Just as worlds at which Timna$_O$ attempts to kill someone who turns out to not be origination-identical with herself are not the worlds we are concerned with (as long as the successful killing would not itself *cause* Timna$_Y$ to be origination-identical with someone else – a possibility we already considered), similarly, worlds at which Timna$_O$ happens to arrive at t0 ex nihilo (say) are not of concern. Since we aren't holding post-t0 facts fixed, we need to allow that Timna$_O$ *can* arrive at t0 ex nihilo. As long as she will not *always* do so, however – that is, as long as she will sometimes arrive at t0 by time travel (we are starting with the assumption that time travel is possible, after all), this won't rid us of the troublesome cases.

But those are troublesome, my opponent may say, because they are the ones where she is a time traveler and so must have survived any earlier attempt on her life: you are holding her survival fixed! To see why this reply doesn't work here, it is important to keep in mind exactly why we are trying to hold only pre-t0 facts fixed: we are aiming to determine whether Timna$_O$'s failure to kill Timna$_Y$ is contingent in a way that makes the failure explicable. And adding in alternative ways for Timna$_O$ to arrive at t0 does not do this, because, for the reasons we've already seen, by selecting for cases in which she time travels, we are still *not* selecting for cases in which she *contingently happens* to fail (and so can time travel for that reason). Unless by happy chance Timna$_O$ appears at t0 ex nihilo,

[5]If my interlocuter remains unconvinced and thinks that a possibility like Timna$_O$ appearing at t0 ex nihilo is what solves the inexplicability worry, I'm happy for her to take my argument as a defense of a weaker conclusion that backwards time travel is only possible if (and at worlds at which) appearing ex nihilo as a fully-grown human – in particular, one who is the older counterpart of a deceased younger individual – is. As I'll argue, I don't think such a possibility does defuse the worry, but even if it does, it seems to me that this weaker conclusion would be sufficiently surprising and interesting in its own right.

Timna$_Y$'s survival is antecedently required – *not* because it contingently happens to occur and we are just holding that fixed (that would be explicable!), but rather, because her survival is an enabling condition even of successful killing. In other words, even a successful attempt requires failure. Since, on pain of contradiction, she can't *both* succeed and fail, she has to fail. [6]

When I talk about it being *antecedently* impossible for Timna$_O$ to succeed, I mean antecedently *conceptually* – the impossibility is conceptually prior to the token failure. It may be useful to think of it in terms of the direction of the in-virtue-of relation. It would be perfectly explicable, as *Telephone* and *Bicycle* are, if the reason that Timna$_O$ had to fail was (i) she contingently happened to, and (ii) we are holding that fixed. That is, it'd be explicable if the necessity of the failure was in virtue of us holding fixed a contingent failure. But that's not the situation, here. Timna$_O$ did not happen to fail, since her even being in position to succeed would first require failure, except in the cases where she happens to appear ex nihilo. (And again, it is not open to my interlocutor to reply with, well, if she didn't appear ex nihilo and didn't fail in the attempt, she just wouldn't be there at all! Neither a successful nor an unsuccessful attempt could explain her absence.) The failure had to occur from the (*conceptual*) beginning, rather than being in virtue of causal happenstance. And if we can't attribute the presence of the foiling mechanism to mere chance, then, that it will be there when needed, i.e., in most cases when Timna$_O$ decides to try to kill Timna$_Y$, remains inexplicable. The mere possibility of Timna$_O$ appearing at t0 in ways other than by time travel makes no difference to this: it is the possibility of backwards time travel itself that gives rise to the problem. [7]

---

[6] Note that this is *not* the case if it is anyone *else* trying to kill Timna$_Y$. If it is Tom, rather than Timna$_O$, trying to kill Timna$_Y$, then success is possible from a pre-t0 perspective, since failure is not an enabling condition for that success.

[7] An anonymous referee brought up another possibility I had not considered. Could not the younger person who is origination-identical with Timna$_O$, or who is at least a candidate for being such, come from the *future* rather than from the past? In that case it would not be logically impossible for Timna$_O$ to kill Timna$_Y$, nor even for her to kill everyone alive at t1: if she succeeded, this would just mean that Timna$_O$'s earlier self came from the future (assuming that Timna$_O$ did not appear at t0 ex nihilo). I see a few problems with this as a potential solution. First, it is not clear to me that it even counts as backwards time travel if Timna's own personal timeline goes from the future-to-past direction. Her traveling to t0 is then just moving in the same temporal direction from which she came. We may decide that this doesn't count as backwards time travel, and set the scenario aside.

Second, putting the first reply aside, it needs to be the case that the person Timna$_O$ kills at t1 was, at least immediately prior to her death, *also* a candidate for being the younger Timna. Otherwise, the

Let me be clear that I *agree* with Lewis and Ismael that, like in *Telephone* and *Bicycle*, in *Autoinfanticide* the failure is already built into the description of the class of cases considered. That Timna_O is alive and able to time travel to t0 entails that Timna_Y survived to adulthood. Crucially, however, *unlike Telephone* and *Bicycle*, in *Autoinfanticide* the failure that is built in *is itself* an inexplicable failure. It is not merely built in that Timna_O *does* fail, the way that it is built in that Ismael and Ernie do. It is built in that she *has to* fail, because, in most circumstances, failure is an enabling condition even of her success. Since it is built in that something *must* go wrong, and not merely that something *does*, we cannot defuse the question, *what will enforce the presence of some foiling mechanism should time-traveling Timna choose to attempt autoinfanticide?*, in the way that analogous questions regarding the other failures can be defused.

Let us take stock. So far I have argued that if *Auto-Infanticide* is explicable, it is *not* for the reason that *Telephone*, *Bicycle* or *Light Switch* are. It is not the case that time-traveling Timna_O is at t0 and in position to kill Timna_Y only because she contingently *happens* to be caused to fail at t1. Instead, the situation seems to be this. If Timna_O is a time traveler and tries to kill Timna_Y then she

---

scenario is not one of attempted *self*-killing. But now things get complicated. Is Timna_O's personal causal history at the moment she arrives at t0 from the future or is it from the past? Since there is no branching time, it cannot be both. If her younger self comes from the future, then the young person at t0 is not the real Timna_Y - i.e., the one who is origination-identical to Timna_O. And if Timna_O's younger self comes from the past, then the potential solution does not apply. So just before t0 it must be the case that Timna_O can be origination-identical with two different younger people, one of whom is in the past (i.e., the person we've been calling "Timna_Y"), and one who is in the future. The "can" here is not merely an epistemic can: it must be metaphysically indeterminate, shortly before t1, whether Timna_O is origination-identical with a younger person from the future or with one from the past. That means, it must be indeterminate from where her causal history lies. Indeed, she must have no determinate causal history. From where do her memories come from, then? The past of t1? The future of t1? Is it metaphysically possible, without branching time, for an individual at a time to have no determinate causal history? It is hard to conceive of how.

Even if it is metaphysically possible for it to be indeterminate, at a time, from which temporal direction an individual's causal history up to that time comes from, we can put such cases aside in the way that we have put to the side cases where Timna_O appears at t0 ex nihilo. All we need to get the problem off the ground is for there to be some possible worlds at which backwards time travel is possible (at those worlds), but at which at least some individuals are not such that the direction of their causal history is indeterminate (or else, who do not have younger individuals who are candidates to be origination-identical to themselves both in the past and in the future). If backwards time travel is possible in these cases, the problem will remain. (Just as before, by putting these cases to the side and focusing on ordinary time-travel scenarios, we are still not selectively attending to cases where Timna_O contingently happens to fail.)

is in a causal loop: her failure to kill Timna$_Y$ at t1 is a cause of her living long enough to get in a time machine at t2, which in turn is a cause of her arriving at t0 and being in position to try to kill Timna$_Y$ at t1. If this causal loop gets started at all, Timna$_O$ has to fail. What's more, it is *not* the case that her happening to fail makes the loop start: once she is in the position to make the attempt, the loop has already started and there is no possibility of success. Nor is it the case that some other event causally upstream from her failure happens to occur which itself causally necessitates the failure. Other key events on the causal chain include Timna$_O$ getting in her time machine and arriving in the past at t0. Although these events are on the causal chain leading to the failure, they do not (causally) necessitate it. Indeed, they are *causally* compatible with her success. It is the additional event, the slip on the banana peel or the jamming of the gun, that makes Timna$_O$ fail rather than succeed. And so far we have found nothing to make explicable the apparently inexplicable fact that an event of this kind will somehow be caused to occur every time a time traveler travels back in time and attempts to kill her own earlier self.

6. Contradiction as Explanation?

I have argued that the Lewis-Ismael-style reply to the inexplicability objection is unsuccessful. But there is another kind of reply to the objection to consider. Perhaps it need not be the case that the failure merely happens to occur by contingent causal happenstance for the failure to be explicable. Perhaps it is enough for explicability simply that there are no worlds at which the time traveler tries to kill her younger self and succeeds. There may be worlds at which there are causal loops involving time travelers trying and failing to kill their infant selves, and worlds at which people appear in the past in ways other than by time travel and then kill their infant selves. But we can all agree that, given our starting assumption that time is one-dimensional, there are no worlds at which time travelers travel to the past and kill their younger selves. So, the defender of time traveler might say, since it follows that every world at which a time traveler travels back in time and tries to kill her younger self is a world at which she will fail, nothing more needs to be said by way of explication.

Nicholas J. J. Smith (1997) pursues this strategy for deflating the apparent inexplicability. He does so by distinguishing explicable from inexplicable kinds of failures in general, and then arguing that the time traveler's failure to kill her younger self falls into the explicable category. To illustrate the difference between explicable and inexplicable failures Smith offers the following examples:

failing to enter your kitchen – "when you try to do so it's as if you run into a glass wall" – despite the fact that your neighbor with an identical floor plan has no problem entering hers (inexplicable) vs. forgetting you have no attic and then always failing to enter the (nonexistent) attic for what appears to you to be some strange conspiracy to prevent you from entering your attic (explicable); searching for an existing fountain of youth but something or other always stops you from finding it despite generally having no problem finding other things you look for (inexplicable) vs. there is no fountain of youth and every attempt made to find it ends in failure for some reason or another (explicable); a barber sets out to shave someone and continuously fails for various reasons (inexplicable) vs. a barber in Newcastle fails each time he sets out to shave all and only the barbers in Newcastle who do not shave themselves (explicable).

According to Smith, the key difference between the explicable and inexplicable failures is the following:

> In both types, one invariably fails to enact a scenario satisfying a certain description (e.g., "entering the attic," "finding the fountain of youth," "shaving all and only those who do not shave themselves"). In one type of case – the one where failure demands an explanation – there are contextually relevant possible scenarios satisfying the description. In the other type of case, there are not…Hence, there is no question of explaining why "such scenarios" are not actualized. Beyond pointing out that there simply are no such scenarios – so whatever happens, one of "them" will not be actualized – there is nothing more to say by way of explanation of the barber's failure. (1997: 160)

It is easy to see how this is supposed to defuse the inexplicability worry. The time traveler's failure to kill her infant self falls into the *explicable* category: "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 traveler commits autoinfanticide…*Whatever happens*, it won't be autoinfanticide because *no scenario at all* satisfies that description." (1997: 161-162)

The problem with this reply is that what makes the scenario seem so inexplicable is that something apparently has to be in place to *cause* the time traveler to fail (she slips on a banana peel, etc.). This differs from Smith's examples, and from other examples involving logical impossibility. In general, logical impossibility is *contained* in itself: it does not work through things or events that are external to it. For example, the reason that the barber cannot shave all and only those barbers who do not shave themselves is that as soon as he shaves all he has not shaven only, but if he shaves only he does not shave all. Nothing needs to stop the barber from succeeding. He need not trip on a banana peel, or get worn out, or forget to bring his razor.

Or consider the painter who desires to make a chair simultaneously red and green all over. If she tries, she will fail. She will fail because for it to be red all over is already, by itself, for it to not be green all over, and vice versa. Her failure will not be due to her slipping on a banana peel or carrying with her the wrong paints. (In my view it does not even make sense to speak as though the cause of her failure is an event of this type.) This is why the time traveler's failure is inexplicable in the way that the barber's or the painter's isn't. Something must stop the time traveler from committing auto infanticide. And logic does not "work through" causation to enforce itself in this way. That is just not the kind of thing that logic is.

What *is* explicable, in Smith's sense, is that the time traveler is not both alive and dead at any one time. As soon as one is dead, one is not alive, and vice versa. But that is as far as the explicability goes. For something to have to happen to *cause* it to be the case that she is alive rather than dead at t1 (even if being dead at t1 would result in a logical contradiction) remains inexplicable.

We can see the problem in a different way. Let us consider one last example. It has been proven to be impossible to trisect an arbitrary angle using only a compass and a straight-edge.[8] Sam is attempting to do just that. Since the task is impossible, Sam will fail. As with the impossibilities discussed before, what is impossible is some particular *combination* of states of affairs. It is impossible for Timna to be both alive and not alive at a time. It is impossible for a chair to be both red all over and green all over at a time. And it is impossible for Sam to use only a compass and straight-edge, and to successfully trisect the angle. There is no possible world at which Sam succeeds by using only these tools. But the fact that there is no such world cannot explain *which* of the set of mutually inconsistent states of affairs fails to occur at a given world. Which does must have a causal explanation. Another way to put this is that the *logical* explanandum is disjunctive: it is required that either Sam fail in his attempt or that he come up with additional tools to work with (say). Now, in this example, like in the previous ones, there is nothing problematic about supposing that which disjunct obtains is explained by ordinary causal happenstance. Indeed, even if it is part of the description of the scenario that Sam has only a compass and a straight-edge, we have no problem assuming that there is a contingent causal explanation for why he has only these tools. And, *given*

[8] I thank an anonymous referee for this example.

that he does, there is a logical/mathematical explanation for why he will fail in his attempt. So the scenario is explicable.

In *Auto-Infanticide*, in contrast, not only is it the case that Timna$_Y$ cannot be both dead and alive at t1 (explicable); it is also the case that she *must* be alive rather than dead at that time. But logic is not in the business of accounting for *which* of the inconsistent states she is in. The absence of possible worlds at which she is both alive and not alive cannot explain why she is alive at a world, on its own. This is problematic when conjoined with the fact that Timna$_O$ *must* fail even when we don't hold post-t0 facts fixed. [9] That is, it is not because Timna$_Y$ *happens* to survive past t1 that Timnao must fail – in the way that Sam must fail only because he happens to have just one compass and a straight-edge. We saw that there is nothing contingent about Timna$_Y$'s survival, even from a pre-t0 perspective. Since (i) the absence of possible worlds at which a given set of mutually inconsistent states of affairs obtains cannot explain why one of those states of affairs is actualized rather than the other, and since (ii) causal happenstance cannot explain it in this case, either, Smith's proposal fails to explicate how something will always be in place to stop the time traveler. [10]

6. *Auto-Infanticide* and Program and Process Explanations

Let us try to be more precise about exactly what it is that is still in need of explication. Consider the following questions:

Q1: Why *must* Timna$_O$ fail?

Q2: Why *does* Timna$_O$ fail?

In a recent article Baron and Colyvan (2016) argue that by attending to the distinction between *program* and *process* explanations sometimes made in the literature about extra-mathematical

[9] If it is preferred, instead of talking about Timna$_Y$, we can talk about the set of individuals who are candidates for being numerically identical with Timna$_O$ given only facts about Timna$_O$ *at t0*. We can make just the same point using them: they can either all survive or not all survive – logic is not in the business of deciding which of these two states occurs. And yet, it is not contingent which happens, either: at least one *must* survive. By speaking about the set who are candidates for being identical with Timna$_O$ instead of about Timna$_Y$, we ensure that we don't sneak in post-t0 facts about Timna's identity, regardless of if we hold a metaphysical view about identity according to which origination-identity is grounded in nothing less than the individual's entire temporally-extended worm (thanks to an anonymous referee for pressing me on this).

[10] Of course, exactly how Timna$_Y$'s life is saved may be a contingent matter (does Timna$_O$ slip on a banana peel, does her gun jam?) but that she must be saved, even when we don't hold fixed post-t0 facts, is not.

explanations, we can find a satisfying answer to questions like Q1 and Q2. And if that is so, they say, then there is nothing else in need of explication. Process explanations appeal to the actual causes that resulted in the explanandum in question. In contrast, a "higher-level", non-causal program explanation "abstract[s] away from, or otherwise ignore[s], the causal details…the program explanation can explain why a particular explanandum must be the case (for the appropriate modality), as opposed to why it is the case *de facto*." (2018: 64) For instance, a process explanation for why your attempt to fit the square peg through the round hole as a matter of fact failed would appeal to the particular causal processes involved: e.g., a particular part of the peg collides with a particular part of the board. A program explanation for why you *mu*st fail appeals to the geometric properties of the squareness of the peg and the roundness of the hole, e.g., the roundness of the hole and the squareness of the peg ensures that each time you try to force the peg through the hole some part or other will get in the way. (2018: 65) Similarly, Baron and Colyvan maintain that in the case of the grandfather paradox (or *Auto-Infanticide*) there is a satisfying answer to both Q1 and Q2; and that is all we need for explicability. In any given instance in which Timna_O tries and fails to kill Timna_Y there will be a process (causal) explanation for what as a matter of fact stops Timna_O from killing Timna_Y in that instance: e.g., perhaps the gun jams. There will be, in addition, a program explanation for why Timna_O *must* fail to kill Timna_Y: if Timna_O were to succeed, she would bring about a logical contradiction, which is impossible.

The problem is that Q1, as it is written, is ambiguous. Is the question, why $\text{must}_{[<t0]}$ Timna_O fail, or is it, why $\text{must}_{[\infty]}$ Timna_O fail? The program and process explanations can conjointly explain the second but not the first. We've seen that logic alone gets us only that it is not the case that Timna is both alive and not alive at a time. On its own logic cannot explain Timna being alive rather than dead at a time (or vice versa), nor her surviving (or not surviving) an attempt on her life. For the program explanation to explain why Timna_O must fail in her attempt, we need the additional fact that Timna_Y does as a matter of fact survive. But, as I've argued, Timna_Y's survival does not occur only as a matter of fact. If the attacker is Timna_O, then Timna_Y must survive, even from a pre-t0 perspective. And it is *this* – that there *must* be a banana peel (or equivalent) *even when we don't hold fixed that* Timna_Y *survives the attempt* – that the program and process explanations cannot account for. The program explanation can account for the fact that there will be something or other in place – a

banana peel or a jammed gun – only *given* that Timna_Y survives the attempt. The program explanation must *already* take Timna_Y's survival for granted (otherwise it can tell us only that she cannot both survive and not survive): it cannot explain it. And the process explanation lacks the required modal robustness to explain the fact that the banana peel (or the like) not only happens to be there but *must* be there.

7. Conclusion

We've seen that in *Auto-Infanticide* we have an additional fact – Timna_O $\text{must}_{[<t0]}$ fail – to account for. What makes the scenario inexplicable is that the banana peel or suitable alternative must be caused to be there if Timna_O tries to kill Timna_Y. In particular, it is *not* the case that the banana peel *happens* to be there and that this explains how Timna is able to live long enough to later time travel to t0 (if that were the case, the banana peel's presence would be explicable). If the banana peel (or its alternative) doesn't merely happen to be there, what can ensure that it is there if Timna_O tries to kill Timna_Y? Perhaps the forces of logic must stay Timna_O's hand after all. Alternatively, if we are unprepared to accept that logic can be the kind of thing to enforce itself through causation in this deviant way, we can deny the possibility of backwards time travel in a universe without branching time.[11]

Work Cited

Baron, S. and Colyvan, M. (2016) "Time Enough for Explanation", *The Journal of Philosophy*, Vol. CXIII, No. 2, pp. 61-88.

Ismael, J. (2003) "Closed Causal Loops and the Bilking Argument", *Synthese*, Vol. 136, No. 3, pp. 305-320.

Lewis, D. (1976) "Paradoxes of Time Travel", *American Philosophical Quarterly*, Vol. 13, No. 2, pp. 145-152.

Sider, T. (2002) "Time Travel, Coincidences and Counterfactuals", *Philosophical Studies* 110: 115-138.

Smith, N.J.J. (2017) "I'd Do Anything to Change the Past (But I Can't Do "That")", *American Philosophical Quarterly*, Vol. 54, No.2, pp. 153-168.

[11] For invaluable comments and discussion, I thank Arif Ahmed, Sam Baron, Nikk Effingham, Jenann Ismael, Hugh Mellor, Daniel Muñoz, audiences at the International Association for the Philosophy of Time 6th Annual Conference, Modal Metaphysics: Issues on the (Im)possible VII Conference, the London School of Economics, Birkbeck, University of London, the Moral Sciences Club (University of Cambridge), and several anonymous referees.