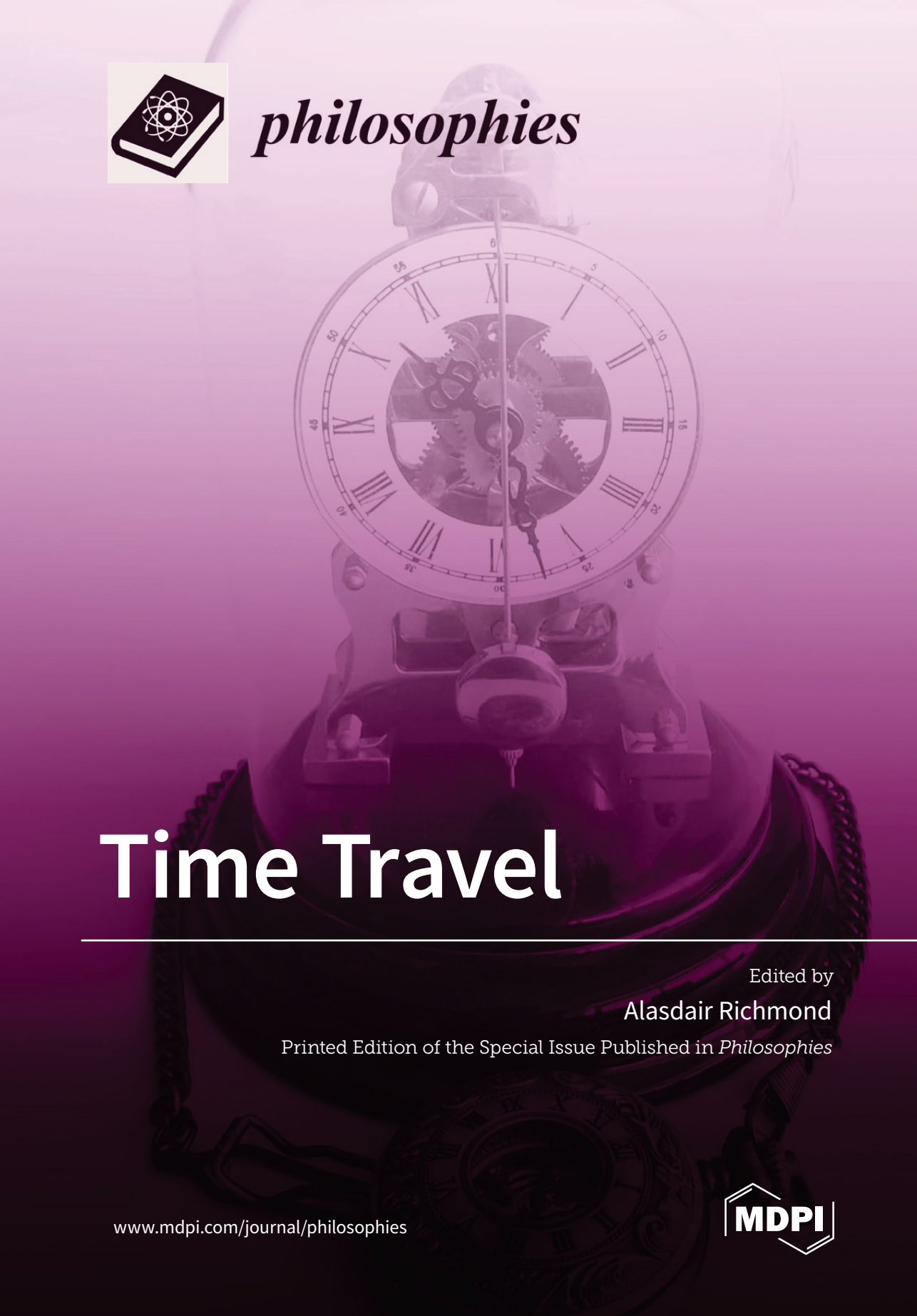
philosophies



Time Travel

Edited by
Alasdair Richmond

Printed Edition of the Special Issue Published in *Philosophies*

www.mdpi.com/journal/philosophies

MDPI

Time Travel

Time Travel

Editor

Alasdair Richmond

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editor
Alasdair Richmond
University of Edinburgh
UK

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Philosophies* (ISSN 2409-9287) (available at: https://www.mdpi.com/journal/philosophies/special_issues/Time_Travel).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-0365-5539-3 (Hbk)
ISBN 978-3-0365-5540-9 (PDF)

Cover image courtesy of Alasdair Richmond

Contents

About the Editor

Alasdair Richmond

Alasdair Richmond is a Senior Lecturer in Philosophy at the University of Edinburgh. He has published on constructive empiricism, the Anthropic Principle, Doomsday arguments, Descartes on immortality, Hume on miracles, time travel and the topology of time. Ever since a sabbatical year in 2008–2009, partly funded by the Arts and Humanities Research Council, his research has centered on the philosophy of time travel in its various guises. His current research covers areas ranging from problems of time traveler freedom, identity and responsibility, through to scientific, computational and even religious aspects of time travel scenarios.

Preface to "Time Travel"

This volume reprints the editor's introduction and all eight contributions to the Philosophies Special Issue on Time Travel. The Special Issue offers a showcase of contemporary work across a broad range of issues connected with the philosophy of time travel. The hope is that the ensuing papers will be of interest to anyone engaged with the philosophies of time and time travel, with metaphysics, and with the philosophy of causation and counterfactuals. The papers are ordered by first appearance and feature work by Kristie Miller, G. C. Goddu, Stephanie Rennick, Nikk Effingham, Richard Hanley, Phil Dowe, Andrew Law and Ryan Wasserman, and Alison Fernandes. The editor wishes to extend thanks and grateful acknowledgements to all the published authors, to everyone who submitted work to the Special Issue, to everyone who carried out refereeing duties for this issue, and to everyone at MDPI who made this publication possible.

Alasdair Richmond
Editor

philosophies

MDPI

Editorial

Introduction to Special Issue *Time Travel*

Alasdair Richmond

School of Philosophy, Psychology and Language Sciences, 40 George Square, Edinburgh EH8 9JX, UK; a.richmond@ed.ac.uk

The philosophy of time travel has an illustrious pedigree, having seen ground-breaking physical and philosophical treatments in the late 1940s and early 1950s from Kurt Gödel. Perhaps the key philosophical work on time travel remains David Lewis's paper 'The Paradoxes of Time Travel' (*American Philosophical Quarterly*, **1976**, *13*, 145–152). As several contributions to this Special Issue attest, virtually all modern philosophy of time travel extensively cites, and is actively engaged in responding to, Lewis (1976). Lewis (1976) makes three principal claims. Firstly, Lewis argued that the traditional 'Grandfather paradox' objections to time travel failed. Secondly, Lewis held that time travellers in the past could possess something like ordinary ability. Finally, Lewis argued that there were no in-principle objections to events being self-causing (i.e., forming causal loops). Other important contributions were made by Hilary Putnam, Paul Horwich, Murray MacBeath, D. H. Mellor, Margarita Levin, Kadri Vihvelin, and others. Two recent highlights were full-length monographs from the Oxford University Press: 2018 seeing Ryan Wasserman's *Paradoxes of Time Travel* and 2020, Nikk Effingham's *Time Travel: Probability and Impossibility*.

The philosophy of time travel is positively burgeoning at present, with more and more areas of the subject being illuminated by discussion of time travel issues. Popular areas of interest include time travel and free will, e.g., exploring how far an agent in the past can retain something like normal abilities. Additionally, popular is the discussion of causal, epistemic, and explanatory problems posed by causal loops—cases where a causal chain folds back into the past so that an event can become amongst its own causes. The metaphysics of time, identity, and laws of nature, plus the epistemology of action, counterfactuals, and deliberation, and even philosophy of religion and philosophy of computation have all yielded interesting time travel discussions. It is a great pleasure to introduce this collection of work by noted scholars in the field. My hope is that this Special Issue serves both as a showcase of new work on time travel and as an introduction to the range of different problems being tackled in this field by distinguished practitioners.

Philosophy of time travel can include discussions of how our attitudes towards the different determinations of time can reflect, or be affected by, our views on the nature of time itself. The paper by Kristie Miller, "What Time-Travel Teaches Us about Future-Bias", looks at the question of how our preferences about events (and the different attitudes we adopt towards events in the past and future) might relate to time-structure. Given that we routinely discount past events (or their impact) relative to that of future events, what might our preferences reveal about the nature of time itself? Miller argues that appeals to temporal structure alone (e.g., whether or not past or future events exist) are not sufficient to explain our preferences. Miller considers other candidates for explaining our preferences—for example, one that appeals to causal salience and another that appeals to an event's location in our *personal* time. ('Personal time' is a notion introduced in Lewis (1976) and denotes time as registered in the traveller's frame of reference, as opposed to 'external time' registered in the world at large.) Interestingly, Miller concludes that neither casual salience nor location in personal time suffice to explain why we tend to discount past events.

As noted above, a key topic in the philosophy of time travel in general (and Lewis 1976 in particular) is the causal loop, the chain of events that allows an event to be among

Citation: Richmond, A. Introduction to Special Issue *Time Travel*. *Philosophies* **2022**, *7*, 100. https://doi.org/10.3390/philosophies7050100

Received: 29 August 2022
Accepted: 31 August 2022
Published: 7 September 2022

its own causes. Where most discussions of causal loops focus on problems with, e.g., information being transferred into the past, Steph Rennick's 'Self-Fulfilling Prophecies' looks at future-based issues with causal loops. In particular, Rennick explores interesting analogies between causal loops generated in time travel cases and causal loops generated via either foreknowledge or knowledge which comes from a perspective outside time altogether (e.g., Divine knowledge). Much of Rennick's focus is on exploring how future-derived knowledge can affect present actions and under what circumstances such knowledge can yield causal loops. Besides considering a wealth of thought-experiments and fictional examples, Rennick also considers how far loops involving foreknowledge compare to other loops in terms of their probability and explicability.

Rather *un*like Lewis (1976), who held that only a single time-dimension would suffice to make time travel possible, many philosophers of time travel have argued that time travel either requires, or would be greatly facilitated by, a metaphysical picture which allowed time itself to possess more than one dimension. The paper by Nikk Effingham, "Exterminous Hypertime" continues Effingham's significant investigations (e.g., in his 2002 book cited above) into using multi-dimensional models of time to explore (and address) problems in time travel. Besides ordinary (linear) time, Effingham's models include different kinds of second temporal dimension, or 'hypertime', which allow changes in past events. Noteworthy features of Effingham's account include: a) the attention paid to how long a change to past events might take to propagate through the second time dimension, and b) the detailed discussion of how well popular metaphysical theories of time might accommodate multiple time dimensions (e.g., whether situated on the tensed/tenseless spectrum, or considered in terms of presentist, eternalist, or growing block theories).

'Changing, Annulling and Otherwising the Past' by G. C. Goddu continues the debate over what multi-dimensional time might mean for the possibility of time travel. Amongst other things, this paper reviews and develops the debate over whether or not a time traveller can change the past, in the sense of taking an event that had obtained in the past and making that event not to have obtained. Goddu reviews theories that allow past-changing via an appeal to extra time-dimensions, orthogonal to 'regular' time, or other kinds of non-standard temporal structure. Additionally surveyed by Goddu is the nature of how events might persist and causes might propagate in the different temporal dimensions. Goddu discusses two senses of changing the past, one in which some past event is made different from what it was and the other in which the past event is made never to have occurred at all, and argues that the former at least is logically possible.

As noted above, Lewis (1976) claimed that Grandfather Paradox arguments fail to show that time travel is impossible and instead at most show that time travellers in the past might face certain constraints on action. 'Lessons from Grandfather' by Andrew Law and Ryan Wasserman develops and expands a classic thought-experiment from Lewis (1976). Lewis imagines a time-traveller called Tim, who travels into the past with a view to assassinating his Grandfather when Grandfather was still a youth, i.e., before Grandfather has become a parent himself. What foils Tim's mission? Law and Wasserman offer two new theories to account for Tim's failure. One account looks to the causal fixity of the past as it relates to an agent's behaviour, while the other puts a causal spin on the principle that no self-undermining act can succeed. Law and Wasserman further explore the implications of their theories of Tim's failure not only for the compatibility of determinism and being able to do otherwise, but also for theories of divine foreknowledge.

If, as Lewis (1976) suggested, travellers in the past might seem to act under constraints, the nature of such a constraint would seem to have implications about which counterfactual conditionals might correctly describe their behaviour. Many theories of counterfactuals either fail to apply to, or explicitly avoiding engaging with, cases of backward time travel. However, Alison Fernandes' paper "Back to the Present: How Not to Use Counterfactuals to Explain Causal Asymmetry" explores the possibility of a general method for evaluating counterfactuals that will work in backwards time travel cases too. Usually, counterfactuals

are assessed by an appeal to holding fixed as much of present actuality as can be maintained while still allowing the antecedent of the counterfactual conditional to hold. However, what sorts of facts should be held fixed in such assessments? Fernandes considers different kinds and locations for 'holding the present fixed', including fixity of distant events in the present. Relatedly, Fernandes considers how counterfactuals relate to causal asymmetry and addresses the problem of how to recover causal asymmetry in a world where physics is apparently temporally symmetrical.

The paper by Phil Dowe, "Does Lewis' Theory of Causation Permit Time Travel?", brings into sharp focus two aspects of David Lewis's work not often linked together, namely Lewis's theory of causation (and specifically how well it allows backwards causation) and Lewis's (1976) theory of time travel. As Lewis himself granted, travel backward in time seems almost bound to require backwards causation, causation whose effect temporally precedes the cause. Dowe offers new reasons for thinking that Lewis' preferred counterfactual theory of causation does not mesh well with the backwards causation that Lewis himself believed was (almost certainly) bound to feature in backward time travel cases. Indeed, Dowe argues Lewis's theory of causation inadvertently and against Lewis's own express intentions, effectively rules out backwards causation a priori or at least, rules out the very kind of backwards causation needed to make backward time travel possible. (*En route*, Dowe critiques other attempts to bring tensions between Lewis on counterfactuals and Lewis on time travel.)

Again, as noted above, Lewis (1976) concluded that travellers in the past could retain something like ordinary, everyday abilities, but this conclusion has proved more controversial than other aspects of Lewis's (1976) case. Richard Hanley's 'Autoinfanticide Is No Biggie: The Reinstatement Reply to Vihvelin' addresses an important challenge to Lewis's (1976) analysis of time travel ability. Lewis (q1976) argues that an agent like Tim is able to carry his (would-be paradoxical) mission relative to some facts about his situation but not others. (E.g. Tim can succeed relative to his being a good shot with a steady hand but not relative to his target being his own Grandfather-to-be.) Kadri Vihvelin (e.g., in her 'Killing Time Again', *The Monist*, **2020**, *103*, 312–327) argues that Tim cannot in any ordinary sense be said to be able to kill Grandfather, because his succeeding cannot take place in any world like ours. Hanley's paper develops a challenge to Vihvelin, based on a class of 'replacement' examples where the traveller's target is killed but is replaced by some suitable 'ontological understudy'. *En route*, Hanley considers different kinds of replacement scenario, drawing on (Lewis' 1976 and others) views of classic personal identity cases like teleportation and fission cases.

Absent of a time machine itself, predicting where philosophy of time travel might go next would be a risky undertaking. However, as the above hopefully makes clear, the philosophy of time travel, as encapsulated in the following papers, draws on a wide and growing variety of important philosophical notions and problems. While often drawing/responding to Lewis (1976), the authors collected here all advance new and fruitful theories of their own.

Funding: This research received no external funding.

Acknowledgments: The editor gratefully acknowledges the contribution to the Special Issue by all authors whose work features therein and especially for the quality of all their submissions. Thanks and acknowledgement are also due for the sterling support offered by MDPI staff.

Conflicts of Interest: The author declares no conflict of interest.

philosophies

MDPI

Article

What Time-Travel Teaches Us about Future-Bias

Kristie Miller

Department of Philosophy, University of Sydney, Sydney, NSW 2006, Australia; kristie.miller@sydney.edu.au

Abstract: Future-biased individuals systematically prefer positively valenced events to be in the future (positive future-bias) and negatively valenced events to be in the past (negative future-bias). The most extreme form of future-bias is absolute future-bias, whereby we completely discount the value of past events when forming our preferences. Various authors have thought that we are absolutely future-biased and that future-bias (absolute or otherwise) is at least rationally permissible. The permissibility of future-bias is often held to be grounded in the structure of the temporal dimension. In this paper I consider several proposals for grounding the permissibility of such preferences and evaluate these in the light of the preferences we would have, and judge that we should have, in various time-travel scenarios. I argue that what we learn by considering these scenarios is that these preferences really have nothing to do with temporal structure. So, if something grounds their permissibility, it is not temporal structure.

Keywords: time travel; future-bias; temporal preferences

Citation: Miller, K. What Time-Travel Teaches Us about Future-Bias. *Philosophies* **2021**, *6*, 38. https://doi.org/10.3390/philosophies6020038

Academic Editor: Alasdair Richmond

Received: 17 March 2021
Accepted: 28 April 2021
Published: 10 May 2021

1. Introduction

Let us say that an agent is *apparently time-biased* with respect to some event[1] just in case they have a preference for where in time that event is located. An agent is *time-biased* if, roughly, their preference for where in time some event is located is sensitive to their representation of its temporal location. So, an agent is time-biased if, with respect to two events E and E* which are equally valuable to the self that experiences them, and adjusting for the subjective probabilities of the events occurring, the agent prefers one event to the other in virtue the temporal locations of the events. So, for instance, suppose that Annie the labradoodle prefers to eat liver cake now, rather than later. That preference might be the result of her believing that the liver cake will get progressively staler throughout the day, and hence that the liver cake later will be intrinsically less valuable to her future self than the liver cake now is to her present self; or it might be the result of her believing that someone else will likely eat the liver cake during the day if she does not eat it now, so she thinks that the probability of receiving the liver cake is higher now, than later; or it might be because she has been told that she will receive a giant turkey drumstick if she prefers to eat the liver cake now, rather than later. If reasons such as these are the sole reasons that Annie has the preference she does, I will say she is *merely apparently time-biased*. She does not discount the value of the later liver-cake *because it is later,* she discounts its value because it is intrinsically less valuable later, or because it is less certain later, or because there is something else of value that she will receive if she chooses to eat the liver cake earlier rather than later.

One kind of time-bias is future-bias. An agent is *apparently future-biased* if she prefers that positively valenced events be located in her future, and negatively valenced events be located in her past. An agent is in fact *positively future-biased* if with respect to two positively valenced events E and E*, where E is located in the future, and E* is located in the past, and where E and E* are equally valuable to the self that experiences them, and adjusting for

[1] In what follows we will talk of the location of events, rather than goods, since we will be particularly interested in the location of certain experiences. But nothing is intended to hang on this.

the subjective probabilities of the events occurring, and holding fixed any other relevant factors (such as being given additional turkey drumsticks) the agent prefers E over E*. An agent is in fact *negatively future-biased* if with respect to two negatively valenced events E and E*, where E is located in the future, and E* is located in the past, and where E and E* are equally valuable to the self that experiences them, and adjusting for the subjective probabilities of the events occurring, and holding fixed any other relevant factors (such as being given additional turkey drumsticks) the agent prefers E* over E. This is to say that, holding fixed all the relevant factors, an agent is future-biased when she prefers positively valenced events be located in the future, and negatively valenced events in the past.

It has been thought that people exhibit future-biased patterns of preferences with regard to hedonic events: sensations, such as pleasures or pains, with some, such as Sullivan [1] (p 58) holding that we are future-biased, that "we assign no value to a merely past painful experience or pleasurable experience." Indeed, many philosophers have thought that not only are we future- biased (absolutely or otherwise) but that this pattern of preferences is rationally permissible (Prior [2], Hare [3,4], Kauppinen [5], Heathwood [6].

That we are future-biased was made especially vivid by Parfit [7]. He asked that we imagine certain situations and then declare our preferences. Here is one such situation. You are to imagine that you wake up in hospital suffering short-term, temporary, memory loss. You know that you have to undergo a painful operation, but you do not know if you are just waking up from the operation or if you are still to have it. You see a nurse approaching and you prepare to ask him whether you already had the operation. Would you prefer to learn that you already had the painful operation, or that you are still to have it? If you share Parfit's preference for learning that you have already had the painful operation, then (holding fixed the subjective probability of the operation occurring, and holding fixed its painfulness, and holding fixed any other relevant factors (such as for instance that the quality of the surgery was the same whenever you had it, that you are not being paid to have surgery on one date rather than another, and so on)) you exhibit *negative hedonic future-bias.*

Now imagine instead that you wake up feeling a little groggy, and for a moment cannot remember whether you are experiencing the after-effects of a truly awesome party, or instead the awesome party is to happen this evening. Would you prefer that the party is still to happen tonight, or that it was last night? If you share Parfit's preference that you learn that the party will be tonight, then (holding fixed the subjective probability of the party occurring, and holding fixed its pleasurability, and holding fixed other relevant factors (such as, for instance, whether there will be a fire on the premises when the party is underway; whether the neighbours will call the police to report a noise complaint; whether the kitchen will be damaged, (and so on)) you exhibit *positive hedonic future-bias.*

More generally, if, holding fixed all relevant factors you have this pattern of preferences, then you attach greater evaluative weight to pleasant and unpleasant experiences when they are in the future than when they are in the past. That is why you prefer negative events to be in the past, and positive events to be in the future.

Parfit [2] (p 173) appears to agree, at least regarding negative events. He writes "I do in fact regard my past suffering with complete indifference. I believe that, in this respect, most other people are like me". Others, such as Suhler and Callender [8] and Yehezkel [9] have thought that we *non-absolutely* discount the past: that we attach some value to past events, but that we discount the value of those events relative to future events.

Recent empirical work supports the contention that people are positively hedonically future-biased: they prefer that, holding fixed all relevant factors, positive hedonic events be located in the future rather than the past. It also shows that people are negatively hedonically future-biased: they prefer that, holding fixed all relevant factors, negative hedonic events be located in the past rather than the future (Caruso, Gilbert and Wilson [10] Greene, Latham, Miller and Norton [11]).

Contemporary work has focused both on whether future-bias is rationally permissible, rationally obligatory, or rationally impermissible. The focus has more frequently been on

the question of whether it is rationally permissible[2]. It is this question to which this paper addresses itself.

Those who think that future-bias is at least rationally permissible often ground its permissibility in features of time itself. This trend started with Prior [2] who pointed out that we have temporally asymmetric attitudes—we are, for instance, relieved that (some) events are past, whereas we dread that (other) events are future—and he suggested that these asymmetric attitudes are both explained by, and rendered rationally permissible by, the structure of time.

In what follows Section 2 I consider a number of proposals for grounding the rational permissibility of future-biased preferences. I then consider these proposals in light of a number of time-travel scenarios. In Section 3, I argue that what time-travel teaches us is that grounding the permissibility of future-biased preferences in absolute pastness and futurity fails. In Section 4 I argue that consideration of time-travel scenarios also shows us that grounding said permissibility in objective pastness and futurity likewise fails. I think the intuitions upon which I draw in these two sections are likely to be fairly robust and general. That is, I think you will share them. So, these two sections show that we have reason to think that if future-biased preferences are rationally permissible, they are not rendered rationally permissible by the structure of time itself. In the following two sections I consider two other proposals for grounding the rational permissibility of future-biased preferences. In each of these sections, I present still more cases, and rely on further intuitions. Section 5 considers the prospects for grounding the rational permissibility of future-bias in *subjective* pastness and futurity. Section 6, considers the prospects for grounding the rational permissibility of future-bias in causal relevance. In each section I present cases, and elicit intuitions, that give us reason to think that neither subjective pastness/futurity nor causal relevance ground the rational permissibility of future-bias. However, here I have to put my hands up. I do not think that all of you will share the intuitions in question. So, what these sections show is that insofar as you share my intuitions in these cases, you have reason to think that we cannot ground the rational permissibility of future-bias in either of these features of the world. Ultimately though, I will happy if you come away from this paper having only been convinced of what I say in Sections 3 and 4.

2. Grounding the Permissibility of Future-Biased Preferences

Let us distinguish the *absolute* past/future from the *objective* past/future and from the *subjective* past/future.

Absolute Past/Future: An event, E, is in the *absolute past* iff E was in the metaphysically privileged present[3]. An event E* is in the *absolute future* iff E* will be in the metaphysically privileged present.

Objective Past/Future: An event, E, is in the *objective past* relative to an agent A located at temporal location T, iff E is earlier than T. An event, E*, is in the *objective future* relative to an agent A located at temporal location T, iff E* is later than T.

Subjective Past/Future: An event, E, is in the *subjective past* of an agent, A, iff E is earlier, in the personal-time of A. An event E* is in the *subjective future* of an agent, A, iff E* is later, in the personal-time of A.

Hence, E is absolutely past only if there is a metaphysically privileged present set of events, and E was metaphysically present, but is no more. To capture this I will often simply talk of E being earlier than the metaphysically privileged present, but this is not intended to imply that E exists[4]. *Mutatis mutandis,* for E* being in the absolute future.

[2] See Prior [2], Hare [3,4]; Kauppinen [5], Heathwood [6] Pearson [12] and Dorsey [13] Brink [14], Greene & Sullivan [15] Sullivan [1] and Dougherty [16,17] Hedden [18].

[3] If E is extended it might be that some of E is still in the metaphysically privileged present even though the rest is absolutely past. We can say that E is completely absolutely past iff E was in the metaphysically privileged present, and is no longer. For most of the purposes of this paper I am interested in events that are completely absolutely past, but for simplicity I will just talk of events being absolutely past.

[4] In case you think that relations like earlier-than are existence entailing, and this characterization is unfriendly to presentists.

So, there are only absolutely past or future events in worlds in which time robustly passes: worlds in which some version of the A-theory is true and hence in which some events are in the metaphysically privileged present, and which events those are, changes.

By contrast, what is *objectively* past, or future, is always relative to a temporal location. So, this characterisation is consistent with a B-theoretic, or block universe, view of time on which past, present, and future events exist and are related by static earlier-than and later-than relations, with no set of events being metaphysically privileged by being objectively present. Given a block universe view, pastness and futurity are not absolute notions: what is past relative to one event is future relative to another[5].

Finally, the *subjective* past (or future) of agent A includes those events that are earlier (or later) in the personal-time of the agent in question. Here, I conceive of personal-time as Lewis [19] does. Roughly, then, personal-time is time as it is measured by the time-traveller. It is the time as measured by the wristwatch of the time-traveller but also, more generally, as measured by the contents of the time machine: the amount of food the traveller digests during the period, the amount of decay of particles in the time machine, and so on. Personal-time is what allows us to make sense of the idea that one can travel backwards in time 300 years, and do so via a journey that takes only 5 minutes. In such a case we have travelled 300 years in time, and have taken 5 minutes of personal-time to do so.

So, I assume that we can order a person's person-stages in terms both of their order in time and in terms of their order in personal-time. Suppose Freddie gets into a time machine and travels back 300 years in time, taking 5 minutes to do so. The person-stage that steps out of the machine will be earlier, in time, than the stage that steps into the machine. However, when we order Freddie's stages in terms of personal-time, we will say that the stage who steps out of the machine is later, in personal-time, than the stage that steps into the machine.

In the remainder of this paper I focus on what we can learn about future-bias by thinking about time-travel. In particular, I will focus on backwards time-travel: time-travel in which one travels to some past time. Perhaps there are also interesting lessons to be learned by considering forwards time-travel. That, however, is a project for another day.

First, then, a brief note on how I am thinking about backwards time-travel (henceforth just time-travel). I will suppose that X travels backwards in time just in case X's departure from one time, t, is a cause of its being true that X arrived at some earlier time, t-minus. So, in order to travel backwards in time it needs to be that the same person arrived at the destination as departs (where 'being the same person' might amount to the there being two person-stages that are stages of the same person), and its being true that the departure event caused it to be the case that X existed at t-minus.

I frame things in this way in order to allow that the destination time might no longer exist at the departure time (if presentism is true, for instance). In such cases a person will count as time-travelling just in case it is, presently, the case that they did exist at some past time, and that their having existed at that past time causally depended on their getting into the machine in the present time[6].

Finally, I assume that time-travel does not involve 'moving the present' in a way that 'undoes' reality. On some views of time-travel (such as that of van Inwagen [20]) in an A-theoretic world, it is not simply the case that a time-traveller always arrives at a time when it is present, but rather, that they take the present with them. I take the difference, here, to be that in the former case a time-traveller travels to a time when it is present (just as they travel to a time when it exists), while in the latter case the time-traveller 'picks up' the present and takes it with her, effectively rewinding time altogether, so that moments that were present but are no longer, cease to be such that they were present at all. So, on a growing block model, for instance, the time-traveller travels back in time and in doing so

[5] Moreover, if you're worried about special relativity, you can suppose that some events are neither objectively past, nor future: these are the ones that relative to L are in the absolute elsewhere. You might think these are neither earlier, nor later, than L. Then you can think that everything in L's backwards light cone is objectively past relative to L, and that everything in L's forward light cone is objectively future relative to L.

[6] Whether this really is time-travel is of course debated.

deletes all of the block back to the moment at which they arrive. The block then re-grows from then onwards. In general, I think, most of what I say holds true regardless of how one is thinking about time-travel. Additionally, I will explicitly consider this view later in the paper. However, it is worth noting that this is not what I have in mind when I talk of time-travel, and perhaps sometimes this makes a difference to what I say.

3. The Absolute Past

Let us begin with a proposal for grounding the rational permissibility of discounting the value of past events, which appeals to those events being absolutely past. This proposal was probably first suggested by Prior [2] who argues that verbal expressions of time-asymmetric attitudes like relief require an irreducibly tensed semantics, which in turn implies (the speaker's belief in) a metaphysics of irreducibly tensed facts or properties[7]. One way to make sense of this idea is to suppose that the time-asymmetric nature of our attitudes grounds the rational permissibility of discounting the past, and, in turn, the rational permissibility of the asymmetry of those attitudes is grounded in there being irreducibly tensed facts. That is, it is rationally permissible to have different attitudes towards past events compared to future ones—regretting or feeling relief about past events, and anticipating or dreading future events—because the former are in the absolute past, and the latter are in the absolute future. In turn, there being these asymmetric attitudes renders it permissible to discount the value of past events relative to future ones.

A slightly different way to argue for the idea that it is rationally permissible to discount the value of absolutely past events goes via the thought that our experience or phenomenology of robust temporal passage[8] grounds the rational permissibility of discounting the past (Schlesinger [21] Craig [22]). I will take it that the phenomenology of robust passage consists in the experience of a change in which events are in the metaphysically privileged present. It might also include the experience of future events coming progressively closer and of past events receding. Then, the idea may be that because events in the absolute past are moving away from us and those in the absolute future are moving towards us, it is rationally permissible to discount the value of the former relative to the latter.

As a matter of fact, neither of these suggestions strikes me as persuasive. First, it's not clear exactly why some events being absolutely past, and some absolutely future, grounds our having asymmetric attitudes towards those events. Second, it's not clear why, even if those attitudes are rationally permissible and are rendered rationally permissible by some events being absolutely past and some absolute future, that this in turn renders it rationally permissible to discount the value of past events. It does not obviously follow from the fact that we *anticipate* some event E when it is future and later feel *relief* that E is past, that we should discount the value of E when it is past, compared to when it is future. Why should the fact that we direct the attitude of relief towards an event mean that the event has less value than when we direct an attitude of anticipation towards it? Perhaps it does, but it certainly is not straightforwardly obvious why this should be. Third, even if we do experience future events as lying ahead of us and moving towards us, and past events as lying behind us and receding away from us, as Yehezkel [9] notes, it's hard to see why this would render it rationally permissible to devalue past events relative to future ones. Similar remarks are made by Parfit [7] (p. 178) and Hare [4], (pp 510–11), among others.

Nevertheless, let us begin with what I call the Unqualified Absolute Past Thesis. I call this the unqualified thesis since on this view it is rationally permissible to absolutely discount the value of past events.

Unqualified Absolute Pastness Thesis (UAPT): It is rationally permissible for A to absolutely discount the value of event E, if E is absolutely past.

[7] Prior's argument has been widely criticised.

[8] I will talk of robust temporal passage to distinguish it from what is sometimes known as anodyne temporal passage, where this latter is consistent with a block universe view of time. Anodyne passage, on this view, is something like the succession of events. By contrast, robust passage consists in the changing of which events are absolutely present.

In what follows I argue that by considering time-travelling scenarios we have good reason to conclude that UAPT is false. To show this, I will describe a series of scenarios and ask you to reflect on your own preferences in these cases, and I will hope that your preferences are roughly like mine. I will also ask you to reflect on which preferences you think are rationally permissible in these cases. That is, I will ask you to reflect on which preferences you think you should have, in these cases.

In doing so, I will suppose that each of us is able to ask ourselves not only what preference we in fact have, but also whether we think those are the preferences that we should have. So, I take it that in principle one could know of oneself that one would in fact prefer the greater pain in the past, to the lesser pain in the future. However, perhaps one also judges that this is not the correct preference to have, precisely because this is to prefer a state of affairs in which, overall, one is worse off.

In the following scenarios I ask that we reflect not only on what we prefer, but also on what preference we think we should have. I assume that our judgements about the rational permissibility of preferences provide some, albeit defeasible, evidence about the rational permissibility of those preferences. So, if we are inclined to judge that we should prefer x to y, then I take it that this is defeasible evidence that it is not rationally permissible to prefer y to x. It is, of course, only defeasible evidence since our intuitions might be mistaken.

Still, defenders of the rational permissibility of future-bias tend to appeal to these kinds of judgements and intuitions in defence of the permissibility of the relevant pattern of preferences. Moreover, they tend to appeal to the idea that it is the structure of time itself that gives rise to these judgements and ultimately vindicates them. So, I take it that such philosophers would find it worrying if, in the scenarios in question, the preferences we have, and think we should have, are not ones that tend to support the idea that future-bias, if rationally permissible, is made permissible by the structure of time. For these philosophers are precisely ones that take seriously the role of our judgements about the rational permissibility of certain preferences, in certain scenarios.

Further, I will take it that our preferences (mine, and yours dear reader) regarding when in time we would prefer some event to be located, provide defeasible evidence of people's preferences more generally in these cases. In an ideal world I would present you with empirical evidence about all of these cases. This world is not ideal. Still, there is evidence that is relevant, and along the way I will point to this. In addition, we have more general evidence about the extent to which in this domain, philosophers' own preferences mirror those of non-philosophers, and the extent to which philosophers' predictions about non-philosophers' preferences are correct.

So, let us quickly consider this general sort of evidence. That evidence suggests that philosophers' own first-personal preferences—that is, their preferences about where *they* would like events that *they themselves* will experience, to be located in time—are the same as non-philosophers' first-personal preferences, and, in turn, their predictions about non-philosophers' preferences (which are made on the basis of their own preferences) turn out to be accurate. Using the sorts of scenarios described by Parfit [7] and Hare [4] philosophers predict that people will be hedonically positively and negatively future-biased. Those predictions are vindicated by empirical research (Greene, Latham, Miller and Norton [11]

There are really only two places where philosophers' predictions about non-philosophers' preferences (in these sorts of cases) have been shown to be inaccurate. The first is with regard to third-personal preferences: these are the preferences that we have over where in time *someone else* experiences some event. Philosophers predicted that third-personal preferences would be unlike first-personal preferences, in that they would fail to exhibit future-bias (Parfit [7] (p 181) Brink [14] (p 378–9), Greene and Sullivan [15] (p 968), and Dougherty [16] (p 3)). For instance, Parfit writes

> "I am an exile from some country, where I have left my widowed mother. Though I am deeply concerned about her, I very seldom get news. I have known for some time that she is fatally ill, and cannot live long. I am now told something new. My mother's illness has become very painful, in a way that drugs cannot relieve.

> For the next few months, before she dies, she faces a terrible ordeal. That she will soon die I already knew. But I am deeply distressed to learn of the suffering that she must endure. A day later I am told that I had been partly misinformed. The facts were right, but not the timing. My mother did have many months of suffering, but she is now dead"

Parfit says that although he is deeply distressed about his mother's suffering, he has no preference regarding whether her suffering has already occurred and is now passed, or whether it is still to come. So, he thinks that in a third-person condition we will be indifferent about the location of hedonic events, in a way that we are not indifferent in a first-person condition.

This prediction is not borne out (Greene, Latham, Miller and Norton [11]). Instead, people's preferences about where in time other people's experience lie mirror their own preferences: that is, they also show future-bias. In this paper I focus only on first-person preferences. So, we can have some confidence that consideration of our preferences (you and I, dear reader) can give us some insight here.

Let us consider our first case.

Case 1:

> You have access to a time machine that is able to take you backwards in time and deliver you to the past. The flight will take 2 h in your personal-time, and you will arrive 300 years into the past. You are not able to work the machine on your own, and you are given several options by 'Time-travellers Journeys', the company that owns the machine and runs time-travel journeys for profit. You can either depart for the past *tomorrow*, or you can depart *the day after tomorrow*. If you depart tomorrow, due to moderately elevated pressure in the time machine (which you will not notice during the flight) when you arrive 300 years in the past, one of the fillings in your tooth will have expanded and you will need emergency dental surgery. You will need to have that surgery in the past, since the time machine takes 24 h to re-power after a journey. That surgery will take 3 h and be very painful, since there were few anaesthetics 300 years ago and they used peddle-drills. So, the surgery will be *very* unpleasant and painful indeed. If you depart for the past the day after tomorrow, the pressure in the time machine will not elevate, and the filling in your tooth will remain where it is. When you arrive 300 years in the past you will not require dental surgery. Would you prefer to take the trip that departs tomorrow, or the trip that departs in two days' time?

Is it rationally permissible to prefer the trip that departs tomorrow, to the trip that departs in two days' time? Well it might be, if certain relevant factors were not held fixed. For instance, suppose that an assassin has been set on your tail and you have reason to believe that if you are around tomorrow the assassin will find you. Then, you would have strong reason to prefer the trip that departs tomorrow, despite the dental surgery involved. For that minimises your chances of being assassinated. Since we are interested in what might ground the rationality of future-biased preferences, not merely apparently future-biased preferences, though, let us hold fixed all these relevant factors.

Holding fixed these factors, my prediction is that both you and I would prefer to take the trip that departs in two days' time. Here is prediction number two: you and I both judge that it is *not* rationally permissible to prefer to take the trip that departs tomorrow. This gives us reason to think that UAPT is false. According to UAPT, since the painful dental procedure is absolutely past, it is permissible for you to absolutely discount its value. So, it is permissible for you to be indifferent between these two trips. That seems wrong.

Now, perhaps you think that backwards time-travel and robust temporal passage are inconsistent. One cannot travel back in time in any world in which time robustly passes.

There are arguments to that conclusion.[9] The best of these, to my mind, appeal to the fact that backwards time-travel requires retrocausation. However, one might argue, the change in which events are present is inextricably connected to the direction of causation. What it is for a set of events to change from being present to being past is for those events to bring it about that a new set of events is present, and they do this via some causal process. If so, the direction of causation must align with the direction of the flow of time. Hence, there can be no backwards causation in worlds with robust passage, and hence no backwards time-travel either. Others think that other considerations are more potent, such as the idea that if presentism is true, then there is nowhere to travel to. At best, one can make it the case that some past-tensed truth is the case: namely that one did exist at some earlier time (Sider [19]). However, one should not think that there is any sense in which one is *about* to travel anywhere (since the anywhere in question does not exist).

Regardless, if one thinks that robust passage is inconsistent with time-travel then two options present themselves. First, it might be that Case 1 is impossible. You will think this if you think that every world with time contains robust passage (i.e., a block universe world is impossible). Then, backwards time-travel is impossible. Alternatively, you might allow that block universes are possible, and hold that backwards time-travel occurs only in these worlds. However, in these worlds there is no absolute past or future. So, Case 1 must be describing a world that lacks an absolute past or future. Hence, it is not a counterexample to UAPT. In either case, Case 1 provides us with no reason to doubt that what grounds the rational permissibility of our discounting past events is that those events are absolutely past. (Of course, if you think that actually, there is no robust passage then UAPT is not going to be appealing since even if it does render it rationally permissible to discount the absolute past, there is actually no absolute past and so it does nothing to explain why our actual discounting of the past is rationally permissible, assuming that it is).

I do not want to adjudicate the issue of whether backwards time-travel is consistent with robust passage. Those who think it is should conclude that Case 1 gives us reason to think that UAPT is false. Indeed, it is worth noting here that if one conceived of passage-friendly time-travel as Van Inwagen [13] does, in terms of picking up and moving the present, then UAPT is especially puzzling. On that view of time-travel one basically 'deletes' all of the past up until the moment to which one travels. So, if it is now the case that Mother Theresa saved, let us say, 1500 lives, if I travel backwards to a time before she begins her good works, I start time from that moment. Then, there will no longer be any truths about those good works, and indeed, time might unfold in such a manner that she no longer performs any such works.

Philosophers have worried that if this were how time-travelled works, then this would raise some pretty hefty ethical issues about time-travelling[10]. For every time someone time-travels they wipe out many lives. However, if UAPT were correct then we should have no such worries. After all, what is being wiped out is all absolutely past. Yet, if backwards time-travel did work this way we surely should worry. So, this suggests that there is something wrong with UAPT.

I think that even those who hold that backwards time-travel is incompatible with robust passage should concede that Case 1 shows us something important. When we think about Case 1, the reason we are inclined to draw the conclusions we do—namely that it is not rationally permissible to prefer to travel tomorrow over the day after—is that the relevant event of the painful dental surgery is in the subjective future of the traveller. Even if this event takes place in the absolute past, as far as the traveller is concerned, they are events towards which one has the same forward-looking attitudes as one does towards events in the absolute future.

[9] For arguments of this kind see Miller [23,24], Sider [25], Slater [26] and Hales [27]. For arguments to the conclusion that time-travel is consistent with robust passage see Monton [28], Daniels [29] Keller and Nelson [30] and Hall [31] and Van Inwagen [13] (though the kind of time-travel that is at issue in van Inwagen's case is rather different).

[10] See for instance Bernstein [32].

Even if, as a matter of fact, the subjective past and absolute past cannot come apart because backwards time-travel is impossible, or is impossible in worlds that contain an absolute past, it seems clear that what is doing the work in this case is the subjective, rather than the absolute, location of these events. This tends to suggest that if something grounds the rational permissibility of discounting past events, it is not that they are absolutely past, but rather, that they are subjectively past.

This, of course, does not show us that UAPT is false. Perhaps what grounds it being rationally permissible to absolutely discount past events is that those events are subjectively past, and what grounds their being subjectively past is that they are absolutely past. Still, there is something uninformative about UAPT. If one thinks that there are worlds that fail to contain robust passage, then one will think that there are other grounds for the rational permissibility of discounting the past. For one will allow that there are worlds in which events are subjectively past, but not absolutely past, and that in those worlds what grounds the rational permissibility of discounting the past is those events being subjectively past.

What we wanted to know was what kind of thing renders discounting the past rationally permissible. In this event it seems right to say that it is that event being subjectively past. It is then a further question what grounds an event being subjectively past. Perhaps sometimes this is grounded by the event being absolutely past, and sometimes not. So, the maximally informative answer to our question—the thing that is really doing the normative heavy lifting—is the status of the event with respect to being subjectively past or future, not its status in being absolutely past or future.

There is also a more general problem with UAPT. Namely, it's very hard to see why the *mere* passage of time should render discounting the past rationally permissible. First, the mere passage of time does not seem to enshrine any sort of *asymmetry* between past and future that would license this normative stance. Presentism and the moving spotlight view, for instance, treat past and future events as sharing the same ontological status. So, that status cannot contribute to its being rationally permissible to differently value those events. While the growing block view holds that past events exist while future ones do not, it's hard to see how that asymmetry could do the work required either. Pre-theoretically, you might have thought that things would go the other way around: that we should more highly value those things that exist, over those that do not!

One suggestion at this point is that what grounds the rational permissibility of our preference for discounting the value of past events is indeed that they are absolutely past, but that this ground goes via the fact that events that are absolutely past have a different intrinsic character to those that are not.

In the last few years various dynamical theories of this kind have been articulated. For instance, some versions of the growing block view are ones on which the intrinsic property of some event E, when E is present, are different from its intrinsic properties when E is absolutely past. One notable view has it that there are no absolutely past phenomenal properties (Forrest [33,34]). Hence, there are no absolutely past pains: there are only events that *were* pains, *when* they were present. There are similar sorts of versions of other dynamical theories, on which absolutely past events/objects are non-concrete (Williamson [35]) or in which they lack ordinary properties such as height, weight, colour, and so on (Cameron [36]).

Suppose that agent A is absolutely present. Now consider some event E that is absolutely past. When E was present, E was painful for A. However, now that E is absolutely past E is not painful, and hence not painful for A. Since for the A-theorist what is true, *simpliciter*, is what is presently true, it follows that E is not painful, *simpliciter*. So, it makes good sense for A to attach no value at all to E.

Notice, though, that if this is how you think things are then you do not think that we exhibit future-biased preferences at all. Instead, you think that we exhibit merely apparently future-biased preferences. After all, if things are this way then everyone, including *time-neutralists*, will agree that we should locate negatively valenced events in the past, and positively valenced events in the future.

Time-neutralists think that we should prefer that arrangement of events that maximises from a time-neutral perspective. That is, our preferences should not be sensitive to our representation of where in time events are located, though of course they should be sensitive to the intrinsic properties of the event, and the subjective probabilities that the event will occur. The time-neutralist will naturally recommend that we attach no value to past events if past events have no value. Or, to put things another way, on this view it is not that we *discount* the value of past events: rather, those events simply have no value, and this is the value we accord them.

So, if our world were like this, it would certainly explain, and indeed render rationally permissible (and surely obligatory) our having *apparently* future-biased preferences. However, UAPT would be doing no work here. Rather, something like the following would be true:

Intrinsic Value Thesis: It is rationally obligatory for A to value events at the value they have when evaluated in the absolute present.

Notice that IVT is not the thesis that we should value events at the value they *had*, when they were absolutely present. Rather, it is the thesis that we should value events at the value those events have, when evaluated at the absolute present.

The time-neutralist will accept this thesis (given that there is an absolute present). Then, if past events have no value, IVT entails that we will, and should be, apparently future-biased.

The point here, though, is that if things are this way then UAPT is false, and it is false because in fact we do not have future-biased preferences at all, and so of course the fact that events are absolutely past/future cannot be what renders those preferences rationally permissible. If UAPT is, as it were, in the running it has to be that we do in fact have future-biased preferences, and so it has to be that the intrinsic value of past events is no different from their intrinsic value when they were present. However, as we have just seen, when we make this assumption UAPT still does not look good.

Indeed, there are other reasons to find UAPT dubious. As a matter of fact, there is reasonable evidence that contrary to the predictions of those like Sullivan [1]) and Parfit [7] we do not absolutely discount past events (Greene, Latham, Miller and Norton [37,38]). While we do value past events less than future events, we do not entirely discount them. Make the past event sufficiently awful, and we will prefer a less awful future event over a more awful past event. This is not what we would expect if UAPT were true.

So, one option is that UAPT was implausibly strong to begin with. Perhaps we should only think that there being some difference between absolute past and absolute future events, makes it rationally permissible to non-absolutely discount past events. Call this the Absolute Pastness Thesis.

Absolute Pastness Thesis: It is rationally permissible for A to discount the value of event E, when E is absolutely past.

If we look to Case 2, however, we can see that this is false.

Case 2:

> You have access to a time machine that is able to take you backwards in time and deliver you to the past. The flight will take 2 hours in your personal-time, and you will arrive 300 years into the past. You are not able to work the machine on your own, and you are given several options by 'Time-travellers Journeys', the company that owns the machine and runs time-travel journeys for profit. You can either depart for the past *tomorrow*, or you can depart *the day after tomorrow*. If you depart tomorrow, when you arrive 300 years in the past, one of the fillings in your tooth will have expanded and you will need emergency dental surgery. You will need to have that surgery in the past, since the time machine takes 24 hours to re-power after a journey. That surgery will take **3 hours** and be **very painful**, since there were few anaesthetics 300 years ago, and they used peddle-drills. So, the surgery will be *very* unpleasant and painful indeed. Fortunately though, 'Time-travellers Journeys' will make sure that

you get back to the time machine for all your after-care needs, and will give you a very effective broad spectrum anti-biotic to prevent infection.

If you depart for the past *the day after tomorrow*, your filling will expand the day *before* you time-travel, and you will have the surgery then (tomorrow). Unfortunately, the one benefit of time-travel is that it makes fillings expand in a much less bad way. So, although if you have the surgery tomorrow you will have better technology, the underlying problem will be significantly worse. So, the surgery will still take **3 hours** and be **very painful.** You can then travel 300 years into the past after the surgery is complete, where you will not require surgery.

You are assured that both surgeries carry the same risk of infection, complications, and future problems, and that both are equally painful.

Would you prefer to take the trip that departs tomorrow, or the trip that departs in two days' time?

Again, let us hold fixed other relevant factors in addition to those specified by the vignette, (such as the risk of infection, etc.). So, for instance, we should imagine that the amount of mental distress is the same whenever you have the surgery, (and so on).

Holding fixed these factors I predict that you, like me, are indifferent between these two options and that you think that we should be indifferent between these two options. However, if it were the case that it is rationally permissible to discount absolutely past events and if we did so discount, then we surely would and should prefer the option on which we have the surgery in the absolute past. For then the disvalue of the absolutely past surgery would be less than the disvalue of the absolutely future surgery. However, this seems wrong. So, we have reason to think that APT is false.

So far, then, we have reason to think that if future-bias is rationally permissible, it is not rendered rationally permissible by time having a particular metaphysical structure: a structure in which some events are absolutely past, and others are absolutely future. That does not mean that the structure of time is not doing the relevant work here though. Perhaps it is not *this* structure that matters. Perhaps instead what matters is that some events are objectively past, and others objectively future. It is this possibility that I consider in the next section.

4. The Objective Past

For each of the two theses we just considered that appeal to absolute pastness, we can instead appeal to objective pastness. Then, we end up with the following two theses:

Unqualified Objective Pastness Thesis (UOPT): It is rationally permissible for A to absolutely discount the value of event E, when E is objectively past relative to A.

Objective Pastness Thesis (OPT): It is rationally permissible for A to non-absolutely discount the value of event E, when E is objectively past relative to A.

The problem with each of these should be clear when we reconsider cases 1 and 2. In Case 1 you should prefer to take the trip that departs in two days' time (holding all relevant factors fixed). If so, we have reason to think that UOPT is false since the painful dental procedure is objectively past, and hence according to UOPT it is rationally permissible to prefer to take the trip that departs tomorrow. Likewise, Case 2 gives us reason to think that OPT is false. You should be indifferent between the two options. However, if OPT were true then it would be rationally permissible to discount objectively past events and hence to prefer the option on which you have the surgery in the objective past.

So, the very same considerations that led us to think that time having a dynamical structure does not ground the rational permissibility of future-bias also give us reason to think that time having an objective direction (there being objectively past, and future, evens) does not grounds its rational permissibility either. In sum, then, this suggests that it is not features of temporal structure itself that ground the rational permissibility of future-bias, assuming something does. It must be something else. What else could it be? In what follows I outline two possibilities and consider their plausibility.

5. The Subjective Past

In this and the following section I consider two additional proposals for grounding the rational permissibility of future-bias: proposals that do not appeal to the structure of time itself. The primary aim of this paper is not to definitively argue that these proposals fail. Rather, I want to draw attention to the fact that matters are not quite as plain sailing as you might have thought. To do this I describe several more scenarios and elicit several more intuitions. I do not think that all of you will share all the intuitions I articulate. If you do, then you, like me, have reason to think that the features of the world I consider do not ground the rational permissibility of future-bias. If you do not share my intuitions, or you are unsure what intuitions to have, then for you these proposals remain live options. Still, I hope that you will at least see that there are some puzzles here that need to be overcome.

The first proposal I consider appeals to subjective past and future. It is easy to motivate this view. Why are our judgements different in Cases 1 and 2 than in ordinary cases such as those described by Parfit? In all three cases the relevant events are *subjectively* future even though sometimes those events are also absolutely or objectively past. That this difference is what matters might seem *prima facie* plausible, insofar as we might think that what matters is whether, from a subjective point of view, negative events are "over and done with" or not. Perhaps, then, it is rationally permissible to discount the value of subjectively past events. Since mostly, objectively past events are also subjectively past, it is easy to see why we might have confused the two. Call the first version of this thesis the Unqualified Subjective Pastness Thesis. On this view, it is rationally permissible to absolutely discount events that are in the subjective past.

Unqualified Subjective Pastness Thesis: It is rationally permissible for A to absolutely discount the value of event E, when E is in A's subjective past.

However, is this right? Consider Case 3.

Case 3:

> You have access to a time machine that is able to take you backwards in time and deliver you to the past. You have a very vivid memory of being a young adult and having dental surgery that went on for about 3 hours and was very painful. Sometimes you just cannot shift the memory: the sound of the drill, the sensation in your mouth, the helplessness of sitting in the dental chair. The time-travel company you are journeying with 'Time-travellers Journeys' offers a new service. It allows you to *change* the past. Although it is, now, true that you did in fact undergo 3 hours of painful dental surgery, you can travel to the past and change what happened. In particular, the company will allow you to take special Nanobytes back to the past. You can deliver these Nanobytes to your younger self by surreptitiously putting them into a beverage drunk by your younger self. The Nanobytes will then fix your younger self's dental problems painlessly. You can either travel back to the time when you were a young adult, put the Nanobytes into your younger self's drink and change the past so that your painful dental procedure never occurs, or you can travel to that time and not put the Nanobytes in your younger self's drink, and hence leave the past as it has always been, as a past in which you underwent the painful dental surgery. You are reassured that if you change the past no other unfortunate or unforeseen consequences will occur. Do you prefer to travel back in time to change your painful dental procedure, or do you prefer to leave the procedure as it is?
>
> Again, remember that we are holding fixed relevant factors.

At this point you might be thinking that changing the past is impossible. Set that thought aside (I will return to it shortly). My prediction is that if you are like me, you will prefer to travel back and change the painful dental procedure.[11]

Moreover, if you are like me you will think that it is not rationally permissible to prefer *not* to change the dental procedure. That is so even though the dental procedure is both

[11] Of course, we don't need to appeal to time-travel here. Suppose that God could change the past. Then we can imagine that God simply asks you whether you want your past dental surgery to be changed or not.

objectively and subjectively past. However, if it is permissible to absolutely discount events in the subjective past, then it is permissible to prefer not to change the dental procedure. So, this gives us reason to think that USPT is false.

At this point you might worry that what is really driving intuitions in Case 3 are the changes that would ensure to your (the protagonist's) subjective present and future. For instance, perhaps by changing the past with regard to my dental procedure I thereby bring it about that I no longer remember those awful events and that I am not longer afraid of the dentist. However, these are benefits that accrue to my current and future selves. So, perhaps my motivation for preferring to change the past is really still being grounded in the value I place on subjectively present and future experiences, rather than on the value I place on subjectively past ones.

To control for this possibility, we can reimagine Case 3.

Case 4 is just like Case 3, except that the time-travel company explains to you that although you can change whether or not you underwent the painful dental surgery, doing so will not change your current memories or affective attitudes. So, you will still have the same apparent memories as of undergoing the painful dental surgery. Indeed, you will still believe that you did undergo the surgery. It is just that this belief will be false, and the memories will be merely apparent. So, nothing about your subjective future or past will be altered if you change whether you underwent the painful surgery. With that in mind, do you still prefer to travel back and change the dental surgery? I am guessing that you do. Moreover, I am guessing that you think that this is the correct preference.

Even so, you might object to Cases 3 and 4 on the following grounds. Changing the past is impossible for precisely the reasons that Lewis [12] outlined. There are those who offer an account of changing the past on which the time we travel back to (and wish to change) is extended along some further dimension—the hypertemporal dimension—so that the original 'part' of the time remains as it ever was, and changing the past consists in making it the case that that time has another part in which things go differently (Goddu [39], Meiland [40]). On such a view times endure across the hypertemporal dimension, so that a time can be one way at one hypertemporal location, and a different way at some other hypertemporal location. Defenders of these models of changing the past insist that this is a perfectly good deserver of the moniker. Others argue that this is not changing the past at all: the original past is still there, all we have done is create some other location at which different things happen, than happen at the original locations (Baron [41]). If you take this latter view, then you will conclude that Cases 3 and 4 describe impossible scenarios.

However, one might think, our intuitions about impossible scenarios are not any kind of evidence about what is, or is not, rationally permissible. So, it would be nice if we could construct a case that is like this one but in which the scenario does not describe changing the past. I think we can. That is case 5.

Case 5:

> You have access to a time machine that is able to take you backwards in time and deliver you to the past. You have a very vivid memory of being a young adult and having dental surgery that went on for about 3 hours and was very painful. Sometimes you just cannot shift the memory: the sound of the drill, the sensation in your mouth, the helplessness of sitting in the dental chair. The time-travel company you are journeying with 'Time-travellers Journeys' offers a new service. It allows clients to travel to the past and insert false memories into their younger selves. You cannot imagine why anyone would buy such a service until your friend Freddie suggests to you that you travel back and insert a false memory as of having a painful 3 hours dental surgery when you were a young adult. You wonder why you would do that. Then, Freddie tells you that if you do, there is every reason to suppose that your current memory is in fact that false memory, and that you never really underwent that very painful surgery. Instead, Freddie suggests, perhaps you should travel back in time and give your younger self a special dentistry pill that releases Nanobytes that fix teeth quickly and painlessly. The time-travel company gives you two options. You

> can travel back without the technology to insert the false memory, and without the Nanobytes, or you can travel back in time with both pieces of technology. What do you choose?

To be clear, Case 5 is intended to be a case in which you can causally effect the past, but cannot change it. You are confident that whether or not you had the past painful dental procedure counterfactually depends on whether you travel back with the Nanobytes and false memory technology. If you travel back with these technologies then this will make it the case that you never had the painful procedure: you just seem to remember having it in virtue of the false memory technology. If you do not travel back with these technologies, then this will make it the case that you had the painful dental procedure. Suppose, too, that whether you in fact had the painful procedure in the past or you merely seem to remember doing so, will make no difference to your subjective future. You will not, for instance, come to fear dentists any less if you come to believe that the memory is merely apparent.

Do you choose to travel backwards with the technology or without it? My intuition in this case is that I would and should travel back with the technology, and I suspect I am not alone here. That is because if I travel back with the technology it is the case that I did not suffer 3 hours of painful dental procedure. Suppose for a moment that you take this view. Notice that the painful dental procedure is not only in the objective past; it is also in the subjective past. However, if it is rationally permissible to absolutely discount the value of subjectively past events, then it is permissible to prefer to travel back *without* the technology. If you think that this is not rationally permissible, then you should conclude that USPT is false.

Moreover, we can construct a similar case that offers a counterexample to the slightly weaker subjective past thesis, which says:

Subjective Pastness Thesis: It is rationally permissible for A to non-absolutely discount the value of event E, when E is subjectively past relative to A.

Here is that case.

Case 6:

> You have signed up with Weird Psychology Tests™ in order to make enough money to afford fetta and avocado toast as well as a mortgage in Sydney. As part of the testing you undergo, you spent yesterday at the testing facility having experiences. You were told that when you awoke this morning you would have temporary amnesia about the nature of those experiences, but that your full memory will return in two days' time. *Tomorrow*, you will be asked whether you want to accept one unit of pain. The experimenters tell you that if you agree to the one unit of pain tomorrow, they will travel backward in time to yesterday and they will make it the case that your experiences yesterday were 6 units of pain. By contrast, if you decline the 1 unit of pain, tomorrow, they will make it the case that yesterday, you received 7 units of pain. So if you choose to decline the 1 unit of pain tomorrow, you will receive 7 units of pain in total, all located yesterday. If you choose to accept the 1 unit of pain tomorrow, you will receive 6 units of pain in total, 5 yesterday and 1 tomorrow. Do you choose to receive the 1 unit of pain tomorrow or not?

If you would choose to have the 1 unit of pain tomorrow, and if you judge that having the alternative preference is not rationally permissible, then you have reason to reject SPT. For tomorrow's pain is in the subjective future. Suppose you discount the 7 pains that are in the subjective past by, for example, 1 unit. Then, you will be comparing 6 units of past pain and no units of future pain, with 5 units of past pain and 1 unit of future pain. In that case it seems entirely rationally permissible to be indifferent between the two outcomes. If you think that you should prefer to accept the one future pain, this suggests that it is not rationally permissible to non-absolutely discount events in the subjective past, and that SPT is false.

Here, I think, intuitions are going to vary. So, this case might not provide you with reason to be suspicious of SPT. After all, my intuitions here are in stark contrast to that of Dorsey [13]. He writes that our inability to affect the past 'makes little difference. One

might simply imagine that it is, in fact, possible to change (causally speaking, that is) the past—by some sort of time-travel machine or the hotline to God. Few would opt for the choice to take the less painful surgery today in order to correct the ten-hour surgery yesterday' [13] (p. 1910). So, I take it that Dorsey, at least, thinks that we will not prefer to accept the one future pain (indeed, he thinks we would not accept a less painful surgery to avoid a 10 h long surgery yesterday).

In fact, there is some empirical evidence we can appeal to here. First, there is evidence about people's preferences in conditions in which the past is *not* causally relevant. Following Parfit's [7] (p 165) famous *My Past or Future Operations* case, Greene, Latham, Miller and Norton [37] used a past-to-future ratio of 10:1 where participants reported whether they would prefer some amount of future pain or *ten times as much* past pain. They found that a majority of people preferred the ten units of past pain (though notably a majority is not everyone: ~18% had the other preference). Interestingly, though, they found the opposite pattern of preference when it came to positively valenced events. When presented with one unit of future pleasure versus ten units of past pleasure, a significant majority of participants (~65%) preferred the *greater pleasure in the past*. In follow up work Greene, Latham, Miller and Norton [38] found that although a majority of people prefer 10 units of past pleasure to 1 unit of future pleasure, they prefer 1 unit of future pleasure to 2 units of past pleasure.

These studies show that although people discount the past, they do not discount it absolutely. Still, they also tend to suggest that perhaps Dorsey is right, and people will not prefer tomorrow's less painful operation over yesterday's more painful one. That suggests that people will not think it rationally impermissible to prefer *not* to have the painful experience tomorrow, contra my prediction.

There is some evidence about this issue too. Latham, Miller, Norton and Tarsney [42] ran a study that presented participants with a vignette that is much like the case described in Case 6. In that vignette people who told that if they chose on additional shock in the future they would have been given 9999 shocks in the past (for a total of 10,000 shocks), but if they choose not to take the additional future shock they should have been given 10,001 shocks in the past, for a total of 10,001 shocks. This study found that the amount of future-bias *diminished* in the condition in which people could choose whether or not to take the additional shock, as opposed to a condition in which they were merely asked their preference regarding where in time the shocks would be located. So, causal relevance does make a difference, contra Dorsey [13].

Still, that study did not find that a majority of people decided to take the additional shock, which is contrary to the prediction I just made about Case 6. There are some features of the vignette used in that study that make me think that people's judgements in Case 6 might be less future-biased. First, the subject of the vignette does not have temporary amnesia; they simply lost track of how many times they were shocked in the past. That plays into the fact that the number of shocks is very high, and one might reasonably think that shocks have a diminishing marginal disvalue: that the difference between 9999 shocks in the past, and 10,001 shocks in the past might be effectively nothing, since perhaps the extra two shocks do not make a noticeable difference (on the back of 9999 shocks). However, if there is no noticeable difference between 9999 and 10,001 past shocks, then since the extra *future* shock will make a noticeable difference, it makes good sense to refuse the extra shock.

I think that a study in which the number of past shocks was much smaller, so that there is an obviously noticeable difference between the number of shocks you get in the past if you accept the additional shock, and the number you get if you do not, would make a difference here. However, of course, this is empirical speculation. What we can say is that if you share my intuitions about Case 6, then you will have reason to reject SPT.

6. Causal Relevance

You might be thinking that all of these cases have something in common. In each case the past event is causally relevant to the preference. In Cases 2 and 3 the past is straightforwardly causally relevant, in that which preference you have determines which choice you make, and your choice is causally efficacious with respect to the relevant past events. The same is true in Case 5. If you prefer that you did not undergo painful surgery, then this preference will lead you to choose to travel back in time with the appropriate technology, and that choice is causally efficacious with respect to the relevant past events. In Case 6 your decision as to whether to accept the single unit of pain tomorrow determines how much pain you experienced yesterday. So, the past event is causally relevant to your decision. This suggests an alternative principle according to which what matters is whether or not an event is causally relevant to your preference.

Causal Relevance Thesis: It is rationally permissible for A to discount the value of event E, when E is causally irrelevant to A.

The rough idea is this. It is permissible to attach less weight to events that are causally irrelevant. The idea is that if we cannot causally affect some event, then nothing we can do will count for or against a choice with respect to that event. This idea was perhaps first suggested by Hume, who writes that the greater effect of future events than past events on the will is easily accounted for. As none of our actions can alter the past, 'tis not strange it shou'd never determine the will' (Hume [43] (Section 2.3.7.6).[12] More recently, Kauppinen [5] has argued that our future-biased preferences are rationally justified by the fact that they have no effect on our choices.[13] Our inability to affect the past also underlies an evolutionary explanation for future-bias suggested by Parfit [7] (p. 186) and Horwich [44] (pp. 194–196) and developed by Maclaurin & Dyke [45] and Suhler & Callender [8].

There is also reason to think this proposal plausible. Greene, Latham, Miller and Norton [46] found evidence that when people are brought to think more agentively about the relevant preference (i.e., to conceive of it more as a choice, rather than a preference) future-biased preferences are significantly decreased (indeed, the study found that people became past-biased under these conditions). That is at least consistent with the idea that future-bias decreases in conditions in which we take ourselves to have choices, where having a choice requires that the relevant events over which we are choosing be causally relevant.

This kind of reasoning would seem to suggest that it is rationally permissible to absolutely discount past events when these events are (as they often are) entirely causally irrelevant to the agent.

The problem is that CRT is not very plausible. Suppose God were to appear to you and tell you that there is some future event E, and that E is special in the following way: it does not matter what happens now, or at any time prior to E, God has made it the case that E will occur regardless. It is not simply that E will happen, and hence that whatever you in fact do, E happens. This is not merely a case of meeting death in Damascus. Rather, it is that, counterfactually, whatever you did, it would have been the case that E happened. So, E is causally irrelevant to your choices: whatever choices you make, these choices make no difference to the occurrence of E.

Suppose, further, that although God knows which event E is, he will tell you only the following: either E is an event that will bring you great pleasure, or E is an event that will

[12] Though this passage is often quoted to associate Hume with the practical irrelevance explanation for future bias, this is probably a mistake: Hume is here talking specifically about effects *on the will*, and the next sentence reads: 'But *with respect to the passions*, the question [of what explains 'the superior effects of the same distance in futurity above that in the past'] is yet entire, and well worth the examining' (Hume [43] 2.3.7.6; emphasis added). In trying to account for the past-future asymmetry with respect to the passions, Hume entertains a number of hypotheses, including a version of the temporal metaphysics hypothesis (2.3.7.9), but does not seem to take the practical irrelevance of the past as an explanation for its weaker effect on the passions.

[13] Kauppinen does not claim that our past-directed preferences are *always* practically inert in the relevant sense. But he holds that when a future-biased preference would influence the agent's choices, or would contradict an earlier preference on which she has already based a choice, future bias is rationally impermissible, *and moreover* is no longer psychologically typical.

bring you great pain. Now, although there is nothing you can do to determine whether E is the event that will bring you great pleasure, or great pain, it still seems clear that you can have a preference regarding which event E is. Namely you prefer that E is an event that brings you great pleasure. Moreover, it does not seem that you will discount the value of E despite it being the case that E is causally irrelevant to your choices.

Suppose you know that the disvalue of E, if E is the painful event, will be minus 200, and you know that the value of E if E is the pleasurable event, be plus 200. There is also some other event, E*. E* is an event that is causally relevant to your choices: you can either bring about E*, or not. E* is worth plus 199. Now suppose that your friend Annie asks you whether you prefer E* over E, conditional on E being positive. I predict that, conditional on E being positive, you prefer E over E*, and that you will judge that any other preference is rationally impermissible. However, if it is rationally permissible for you to discount the value of E because it is causally irrelevant, then the value of E will fall below the value of E*, and you should prefer E*. So, this suggests that we do not think it is rationally permissible to devalue E, even though it is causally irrelevant.

Again though, perhaps you do not share my intuitions here. If so, you do not have reason to think that CRT is false.

7. Conclusions

What have we learned from all this? First, if it is rationally permissible to discount the value of past events then that permissibility does not issue from the structure of time itself. It does not issue from the fact that time robustly passes, (if it does), nor from the fact that some events are objectively past, (or future) relative to others. Perhaps the fact that some event is objectively, or absolutely, past often correlates with some other properties of that event, in virtue of which it is permissible to discount the value of that event. However, if so, it is these other properties that are doing the normative work, not the location of that event in time.

Two candidates that would fall into that category are the causal relevance of an event, and its location in personal time. Generally speaking, past events are both causally inaccessible and lie in the subjective past, while future events are causally accessible and lie in the subjective future. While I think that each of these plays some role in our preferences, if you share my intuitions about these cases, you should conclude that there is reason to think that neither is sufficient to render it rationally permissible to discount the past. That leaves open that these factors, perhaps in conjunction with some other factor(s) do the normative work here, or that there is no normative work to be done, because in fact it is not rationally permissible to discount the value of past events. If, on the other hand you do not share my intuitions in these cases then it might be that you have reason to think that one (or both) of these principles is correct. Perhaps then your task is to try to explain why it is that people like me have contrary intuitions.

Funding: This research was funded by ARC (Australian Research Council) grant number DP180100105 and FT170100262.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Sullivan, M. *Time Biases*; Oxford University Press: Oxford, UK, 2018.
2. Prior, A.N. Thank Goodness That's Over. *Philosophy* **1959**, *34*, 12–17. [CrossRef]
3. Hare, C. Self-Bias, Time-Bias, and the Metaphysics of the Self and Time. *J. Philos.* **2007**, *104*, 350–373. [CrossRef]
4. Hare, C. A Puzzle about Other-Directed Time-Bias. *Australas. J. Philos.* **2008**, *86*, 269–277. [CrossRef]
5. Kauppinen, A. Agency, Experience, and Future Bias. *Thought A J. Philos.* **2018**, *7*, 237–245. [CrossRef]
6. Heathwood, C. Fitting Attitudes and Welfare. *Oxf. Stud. Metaethics* **2008**, *3*, 47–73.
7. Parfit, D. *Reasons and Persons*; Oxford University Press: Oxford, UK, 1984.
8. Suhler, C.; Callender, C. Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry. *Philos. Impr.* **2012**, *12*, 1–16.
9. Yehezkel, G. Theories of Time and the Asymmetry in Human Attitudes. *Ratio* **2014**, *27*, 68–83. [CrossRef]

10. Caruso, E.; Gilbert, D.T.; Wilson, T.D. A Wrinkle in Time: Asymmetric Valuation of Past and Future Events. *Psychol. Sci.* **2008**, *19*, 796–801. [CrossRef]
11. Greene, P.; Latham, A.J.; Miller, K.; Norton, J. Hedonic and non-hedonic bias towards the future. *Australas. J. Philos.* **2020**. [CrossRef]
12. Pearson, O. Appropriate Emotions and the Metaphysics of Time. *Philos. Stud.* **2018**, *175*, 1945–1961. [CrossRef]
13. Dorsey, D. Prudence and Past Selves. *Philos. Stud.* **2018**, *175*, 1901–1925. [CrossRef]
14. Brink, D.O. Prospects for Temporal Neutrality. In *The Oxford Handbook of Philosophy of Time*; Callender, C., Ed.; Oxford University Press: Oxford, UK, 2011; pp. 353–381.
15. Greene, P.; Sullivan, M. Against Time Bias. *Ethics* **2015**, *125*, 947–970. [CrossRef]
16. Dougherty, T. Future-Bias and Practical Reason. *Philos. Impr.* **2015**, *15*, 1–16.
17. Dougherty, T. On Whether to Prefer Pain to Pass. *Ethics* **2011**, *121*, 521–537. [CrossRef]
18. Hedden, B. *Reasons Without Persons: Rationality, Identity, and Time*; Oxford University Press: Oxford, UK, 2015.
19. Lewis, D. The Paradoxes of Time Travel. *Am. Phil. Q.* **1976**, 145–152.
20. van Inwagen, P. Changing the past. In *Oxford Studies in Metaphysics*; Dean, Z., Ed.; Oxford University Press: Oxford, UK, 2010; Volume 5, pp. 3–28.
21. Schlesinger, G. The stillness of time and philosophical equanimity. *Philos. Stud.* **1976**, *30*, 145–159. [CrossRef]
22. Craig, W.L. Tensed time and our differential experience of the past and future. *South. J. Philos.* **1999**, *37*, 515–537. [CrossRef]
23. Miller, K. Time-travel and the open future. *Disputatio* **2005**, *1*, 223–232. [CrossRef]
24. Miller, K. Backwards causation, time, and the open future. *Metaphysica* **2008**, *9*, 173–191. [CrossRef]
25. Slater, M. The necessity of time-travel (on pain of indeterminacy). *Monist* **2005**, *88*, 362–369. [CrossRef]
26. Hales, S.D. No time-travel for presentists. *Logos Epistem* **2010**, *1*, 353–360. [CrossRef]
27. Sider, T. Traveling in A- and B-time. *Monist* **2005**, *88*, 329–335. [CrossRef]
28. Monton, B. Presentists can believe in closed timelike curves. *Analysis* **2003**, *63*, 199–202. [CrossRef]
29. Daniels, P. Back to the Present: Defending Presentist Time Travel. *Disputatio* **2012**, *4*, 469–484. [CrossRef]
30. Keller, S.; Nelson, M. Presentists should believe in time-travel. *Australas. J. Philos.* **2001**, *79*, 333–345. [CrossRef]
31. Hall, T. In Defense of the Compossibility of Presentism and Time-travel. *Logos Epistem.* **2014**, *2*, 141–159. [CrossRef]
32. Bernstein, S. Time Travel and the Movable Present. In *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen*; Keller, J., Ed.; Cornell University Press: Ithaca, NY, USA, 2017; pp. 80–94.
33. Forrest, P. The read but deal past: A reply to Braddon-Mitchell. *Analysis* **2004**, *65*, 358–362. [CrossRef]
34. Forrest, P. Uniform Grounding of Truth and the Growing Block Theory: A Reply to Heathwood. *Analysis* **2006**, *66*, 161–162. [CrossRef]
35. Williamson, T. Necessary Existents. In *Royal Institute of Philosophy Supplement*; O'Hear, A., Ed.; Oxford University Press: Oxford, UK, 2002; pp. 233–252.
36. Cameron, R. *The Moving Spotlight: An Essay on Time and Ontology*; Oxford University Press: Oxford, UK, 2015.
37. Greene, P.; Latham, A.J.; Miller, K.; Norton, J. (ms1) On Preferring that Overall, Things are Worse: Future-Bias and Unequal Payoffs. Available online: https://philpapers.org/rec/GREOPT-2 (accessed on 31 March 2021).
38. Greene, P.; Latham, A.J.; Miller, K.; Norton, J. (ms2) How Much do We Discount Past Pleasures? Available online: https://philpapers.org/rec/GREHMD-4 (accessed on 31 March 2021).
39. Goddu, G.C. Time-travel and Changing the Past: (Or How to Kill Yourself and Live to Tell the Tale). *Ratio* **2003**, *16*, 16–32. [CrossRef]
40. Meiland, J.W. A Two-Dimensional Passage Model of Time for Time-travel. *Philos. Stud.* **1974**, *26*, 153–173. [CrossRef]
41. Baron, S. Back to the Unchanging Past. *Pac. Philos. Q.* **2017**, *98*, 129–147. [CrossRef]
42. Latham, A.J.; Miller, K.J.; Norton, J.; Tarsney, C. Future Bias in Action. *Synthese* **2020**. [CrossRef]
43. Hume, D. *A Treatise of Human Nature*; Oxford University Press: Oxford, UK, 1738.
44. Horwich, P. *Asymmetries in Time: Problems in the Philosophy of Science*; MIT Press: Cambridge, MA, USA, 1987.
45. Maclaurin, J.; Dyke, H. Thank Goodness that's Over': The Evolutionary Story. *Ratio* **2002**, *15*, 276–292. [CrossRef]
46. Greene, P.; Latham, A.J.; Miller, K.; Norton, J. Why are People So Darn Past-Biased? In *Temporal Asymmetries in Philosophy and Psychology*; Hoerl, C., McCormack, T., Fernandes, A., Eds.; Oxford University Press: Oxford, UK, 2021.

philosophies

MDPI

Article

Changing, Annulling and Otherwising the Past

G. C. Goddu

Department of Philosophy, University of Richmond, Richmond, VA 23173, USA; ggoddu@richmond.edu

Abstract: Despite a growing number of models argument for the logical possibility of changing the past there continues to be resistance to and confusion surrounding the possibility of changing the past. In this paper I shall attempt to mitigate the resistance and alleviate at least some of the confusion by distinguishing changing the past from what Richard Hanley calls 'annulling' the past and distinguishing both from what I shall call 'otherwising' the past.

Keywords: time travel; logical possibility; changing; fixing; annulling; otherwising

Citation: Goddu, G.C. Changing, Annulling and Otherwising the Past. *Philosophies* **2021**, *6*, 71. https://doi.org/10.3390/philosophies6030071

Academic Editor: Alasdair Richmond

Received: 3 August 2021
Accepted: 25 August 2021
Published: 30 August 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

> Almost a hundred years ago, science fiction editor, Hugo Gernsback wrote:
>
> "The question in brief is as follows: Can a time traveler, going back in time—whether ten years or ten million years—partake in the life of that time and mingle in with its people; or must he remain suspended in his own time-dimension, a spectator who merely looks on but is powerless to do more?" [1] (p. 610)

His query was in response to several letters challenging earlier stories Gernsback had published in *Amazing Stories*. The letters insisted, that for the time travel stories to be consistent, the time travelers needed to be invisible. (See [2] (pp. 171–173) for discussion of these early 'fan' comments on time travel.)

The underlying concern perhaps, a concern made explicit in later philosophical arguments about time travel (see for example [3] (p. 177)) is that actually travelling to the past would entail changing the past and changing the past is logically impossible, so the best we can do is experience the past via early science fiction's abundant chrono-scopes, chrono-cameras, time-radios, etc. The concern isn't merely that time travelers might step off the safe path and accidentally crush the proverbial butterfly, [4] but that even building the 'safe path' in the first place would 'damage' or 'change' the time line.

Many philosophers resisted these arguments on the grounds that while changing the past is indeed logically impossible, time travel into the past does not entail changing the past—it merely entails affecting the past. [5] So given unrestricted time travel to the past you can visit the building of the pyramids or the Great Wall, you can help the Union or the Confederacy, you can peruse the library at Alexandria, you can do almost anything in the past you might want—you can even try to change the past in some way, say by trying to prevent Booth from killing Lincoln or by trying to prevent the Holocaust. If changing the past is impossible, you will fail, but if you want to try, time travel will certainly allow the trying. (See, for example, [6] for a self-defeating attempt to prevent the Holocaust.)

Even more recently however several arguments have appeared that changing the past is, contra the prevailing view, logically possible. If you want to kill Hitler before 1933 or put Aristotle on a 'better' path, you can, but you will also have to live with the consequences of your changes. (See, for example, [7] for another twist on killing (or not killing) Hitler or [8] for the potential consequences of trying to influence Aristotle.) Despite these arguments resistance and confusion surrounding the possibility of changing the past persists. In this paper I shall attempt to mitigate the resistance and alleviate at least some of the confusion. In Section 2, I first articulate a common way to model the possibility of changing the past and then in Section 2.1 present and reject Nicholas J.J. Smith's [9]

most recent arguments that these sorts of models model avoiding the past rather than changing the past. In Section 2.2, I shall consider the possibility of fixing the past and argue that we must distinguish two types of fixes—one is just a kind of change and is possible, the other is stronger, what I call, following Richard Hanley [10] 'annulling' and is not. In Section 3, I shall further clarify annulling the past by distinguishing strong annulling which is impossible from weak annulling which is a kind of change and is possible. In the process, I shall argue against Sam Baron's [11] conflation of strong annulling and change and claim that Baron ultimately concedes that the sort of changing the past that recent theorists have been interested in is possible. Finally, in Section 4, I shall argue that Peter Vranas' [12] arguments that a certain sort of change, which at first blush sounds like strong annulling, is possible in fact support the possibility of something quite distinct from strong annulment, which I shall call 'otherwising the past'. I shall conclude that despite the high potential for confusion we should be careful to separate the possibility of changing the past from the impossibility of strongly annulling the past.

2. Changing the Past

There are two general strategies in the literature for arguing that changing the past is logically possible. Firstly, one can introduce another temporal dimension or another time-like structure in addition to normal time. (See, for example, [13–18]). Alternatively one can keep just the single temporal dimension, but deny that earlier than/later than are always correlated. (See for example, [19,20], and especially [21]). I shall focus here on the first strategy, though much of what I say below can be adapted to the second.

In some works, such as [13,15], and [17], the second temporal structure is a second orthogonal time dimension. In others [14,22], what we normally think of as time is embedded in another time-like structure, not necessarily orthogonal. Either way, the second temporal structure is generally referred to as 'hypertime', and I shall continue to do so. On either treatment of hypertime, momentary time slices (or the objects or events of those slices) can be hypertemporally extended or occur again such that they have one set of properties, say grandfather being alive, at one hypertime, but grandfather being dead at another.

Let 'u' be a complete universe state at a particular time. Let 't' be times and 'H' be hypertimes. Hence, a universe without time travel could be partially represented as follows in Figure 1:

H_{1901}	…	H_{1921}	…	H_{1941}	…	H_{1961}	…	H_{1981}	…	H_{2001}	…	H_{2021}	…	H_{2041}
t_{1901}	…	t_{1921}	…	t_{1941}	…	t_{1961}	…	t_{1981}	…	t_{2001}	…	t_{2021}	…	t_{2041}
u_{1901}	…	u_{1921}	…	u_{1941}	…	u_{1961}	…	u_{1981}	…	u_{2001}	…	u_{2021}	…	u_{2041}

Figure 1. A hypertemporal universe with no time travel.

Given no time travel has happened, right now (t_{2021}, H_{2021}) Hitler survives past 1921—that is the way the past is right now. But suppose that the first time traveler departs for the past in 2041 (t_{2041}, H_{2041}) and arrives in 1921 (t_{1921}, H_{2042}). Why hypertime 2042? Because on almost all (see [18] for an exception) the hypertime models for changing the past, travelling backwards in time still involves moving forward in hypertime. In 1921 the time traveler kills Hitler and stays in the past to make sure no one else arises to fill the role of Hitler. We could partially represent this universe as follows (making the t's line up) in Figure 2:

In u_{1921} at t_{1921}, H_{1921} no time traveler appears and Hitler is not killed, but in $u_{1921'}$ at t_{1921}, H_{2042} a time traveler appears and Hitler is killed and so all the subsequent universe states change as a consequence.

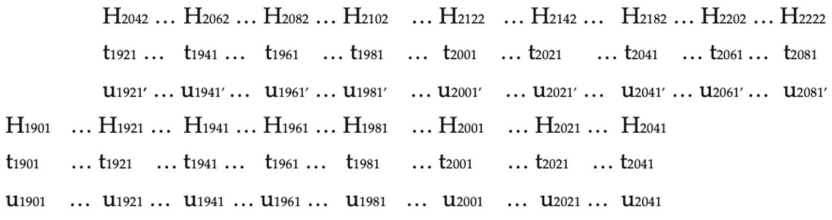


Figure 2. A hypertemporal universe with one instance of time travel.

Suppose no other time travel occurs. What is the past like in 2021 after the time travel? In 2021 (t_{2021}, H_{2142}), the past no longer contains Hitler surviving past 1921, whereas in pre-time travel 2021 (t_{2021}, H_{2021}) he does. In other words, the past used to have Hitler surviving past 1921 (at say H_{2021}), but no longer does (at H_{2042} and forward).

There is a reason I have ordered the representation via the times rather than the hypertimes. Imagine each Htu layer as a layer of paint. I start painting a surface red. I stop and then partway into the red portion start painting over the red with blue and then beyond where I stopped painting red. What does the surface look like? It has a red portion and then a blue portion (some of which overpaints an old red portion). Similarly, the covered up Htu layer is the way things used to be and any uncovered Htu layer is the way things currently are. Hence, a historian writing an accurate history in $u_{2021'}$ would describe events in u_{1901} leading to $u_{1921'}$ (along with the strange appearance of an individual in a strange machine out of thin air and the death of Hitler) leading to $u_{1941'}$ and so on up to $u_{2021'}$. A historian writing an accurate history in u_{2021} would describe events in u_{1901} leading to u_{1921} in which Hitler survives leading to u_{1941} and so on up to u_{2021}. Given that these accurate histories describe the past as it is at a given hypertime and the accurate histories are different, then the past can change. On hypertemporal models times (or the universe states of particular times) happen again at later hypertimes. Hence, t_{1921} occurs at both H_{1921} and H_{2042}. It is that occurring again that allows for change—Hitler not dying in 1921 (t_{1921}, H_{1921}) and then subsequently dying in 1921 (t_{1921}, H_{2042}).

2.1. Avoiding the Past

Despite the growing number of models purporting to model changing the past, some still argue that changing the past is impossible. According to Nicholas Smith [23]: "If there is no bifurcation, of time or place, then there can only be contradiction, not change. Yet even if there is such bifurcation, still there can be no change, only *avoidance*."(emphasis in original). More recently, Smith [9] (p. 690), in defense of his avoidance charge claims that these models are just assuming (or stipulating) that t_{1921} at H_{1921} and t_{1921} at H_{2042} "are one and the same normal time." He goes on: "But this cannot simply be stipulated. The whole point of Task Two[1] is that we need a substantive account of what makes it the case that "1928" at hypertime *b* is the same time as "1928" at hypertime a . . . ".

While it is not fully clear what Smith means by a 'substantive' account he does offer as examples substantive accounts of object identity through time such as endurantist accounts of objects or perdurantist accounts of objects. Interestingly enough, Meiland has been interpreted as an endurantist about the past, i.e., that the past itself is a continuant and I have been interpreted as a perdurantist about the past, i.e., that just as objects perdure by having different temporal parts, temporal moments (or the universe slice at a time) perdure by having different hypertemporal parts. (See Smith, [9]) Van Inwagen [15] can be interpreted as a hypertemporal presentist with a temporal growing block where the whole block is a continuant. Bernstein [17] considers various ontological possibilities for a moving spotlight theory. Indeed, given all these ontological options, I have been deliberately as neutral as possible concerning the underlying ontology, since regardless of the temporal ontology, the crucial piece is the ability to say that at one hypertime the events of time t

were x, but at some later hypertime the events of time t were something different than x, say y.

Admittedly, beyond asking us, say, to imagine that momentary time slices of the universe or the momentary parts of objects there-in have hypertemporal parts in the same way that extended objects can have temporal parts these authors are not fully explicit on hypertemporal identity conditions. Smith, however, seems to accept traditional accounts of object identity through time as candidate substantive accounts. Hence, isn't saying that the hypertemporal identity conditions are just the temporal identity conditions, but applied to objects through hypertime, substantive enough? Take your preferred endurantist account of how object *a* endures from t1 to t2 and apply it to universe states across hypertime. Alternatively, take your preferred perdurantist account of how object *a* perdures from t1 to t2 in virtue of having properly related temporal parts at t1 and t2 and apply it to temporal slices at hypertimes—as long as u_{1921} and $u_{1921'}$ are so related, we can say they are different hypertemporal parts of the time slice that is t_{1921}.

Given that we are concerned with logical possibility and impossibility, the defenders of hypertemporal models might claim that the burden of proof now rests on the objector—assuming that at least some endurantist or perdurantist accounts are themselves logically possible, prove that extending that account to objects enduring or perduring through a second time-like structure would generate an impossibility, otherwise the models stand. Smith, however, concludes his article with reasons to think providing a substantive account for diahyperchronic identity conditions is not possible. Given that diachronic identity conditions rely on some sort of causal dependence, and diahyperchronic identity is supposed to be just like diachronic identity, then the diahyperchronic identity conditions will too. But now Smith objects that we will run into an exclusion argument. He writes: "if we have a full story about how normal time t is (at hypertime b) in terms of how things were earlier in normal time (at hypertime b) then there is no room for a story about how things are at normal time t (at hypertime b) depending on how things are at normal time s at hypertime a." [9] (p. 692)

On the one hand, I am not sure why this is a problem. We are interested in the logical possibility of changing the past—as long as overdetermination itself is not logically impossible then, even if hypertemporal models involve causal overdetermination, these models would still provide logically possible accounts of changing the past. Indeed, given that on these models hypertemporal change only comes about because of time travel, in models such as Meiland's or van Inwagen's in which we can have hypertemporal extension without time travel we would fully expect causal overdetermination. t_1H_1 makes t_2H_1 the way it is because of the temporal causal relation between t_1 and t_2. t_1H_1 also makes t_1H_2 the way it is because of the hypertemporal causal relation between H_1 and H_2. t_2H_2 is the way it is (temporally) because of t_1H_2 and (hypertemporally) because of t_2H_1. Since there is no time travel there is no hypertemporal change and unsurprisingly then, t_2H_2 looks just like t_2H_1.

In my model, the time travel itself makes the universe slices become hypertemporally extended, so we cannot have complete causal redundancy. Consider $u_{1921'}$. It is the way it is (hypertemporally) because of u_{1921} and because of u_{2041} (because of the time travel). $u_{1941'}$ is the way it is (hypertemporally) because of u_{1941} and (temporally) because of $u_{1921'}$ (and the intervening times). Clearly u_{1941} does not suffice to account for all of $u_{1941'}$ since it is the changes in $u_{1921'}$ that accounts for the changes in $u_{1941'}$. But the worry might be that u_{1941} isn't offering any hypertemporal causal input into $u_{1941'}$ at all. To avoid Smith's charge that we are not talking about one and the same 1941, the defenders of the models have to hold that there is indeed a hypertemporal causal connection (or whatever ultimately is the required connection) between u_{1941} and $u_{1941'}$.

Again, I see no problem here. On the one hand, u_{1941} might be offering no distinct causal input other than what $u_{1921'}$ (and subsequent times) is offering. But again, as long as overdetermination is not impossible, it is not at all clear why $u_{1941'}$s contribution, whatever it is, cannot overlap completely with some part of what $u_{1921'}$ (and subsequent times) is

contributing. On the other hand, perhaps u_{1941} is offering distinct causal input. Perhaps what we normally think of as indeterministic quantum events occur the same way they do in $u_{1941'}$ not because of anything about $u_{1940'}$ but because of how those events occurred in u_{1941}. In other words, while temporal causal indeterminacies would make it impossible to predict whether a certain atom present in $u_{1940'}$ would decay in $u_{1941'}$, those same indeterminacies might not be present for hypertemporal causation. If the atom decays in u_{1941} (and is outside the cone of any changes wrought by time travel), then it will also decay in $u_{1941'}$. In this case, even if there was significant causal overlap, $u_{1940'}$ and u_{1941} would still also offer different causal contributions to $u_{1941'}$. Either way Smith's exclusion argument fails.

Smith, however, also offers a slightly different exclusion argument. He writes:

> "Meiland and Goddu want a model in which effects propogate through normal time in the usual way and objects persist through normal time in the usual way. But this is incompatible with the kinds of view one would need to have about causal dependency to think that the same time can persist across hypertime. The former requires thinking: how t_1 is at hypertime b depends on how earlier normal times are at hypertime b and how earlier normal times are at hypertime b suffices for how t_1 is at hypertime b. The latter requires thinking: how t_1 is at hypertime b depends on how t_1 is at hypertime a and how t_1 is at hypertime a suffices for how t_1 is at hypertime b. These claims of causal dependence and sufficiency cannot all be true." [9] (p. 692)

I agree that both dependency/sufficiency claims cannot be true at the same time, as long as we read 'depends' as 'depends only', or else we are right back into the possibility of overdetermination. But since the first claim is incompatible with time travel itself and Smith accepts the logical possibility of time travel, we should just reject the first dependency/sufficiency claim. For example, even without a second temporal dimension, with time travel into the past it is just not true that how t_1 is depends on how earlier normal times and how earlier normal times are suffice for how t_1 is. After all, how t_1 is might depend on how later normal times are. Similarly, in hypertemporal models, how t_{1921} at H_{2042} is depends on how t_{1921} at H_{1921} is and how t_{2041} at H_{2041} is and yet t_{2041} is later in normal time.

In addition, I strongly suspect that many hypertemporal theorists will reject the second claim as stated and still hold that one and the same time can be extended across or endure through hypertime. In the example previously given in which u_{1941} and $u_{1940'}$ offer partially overlapping causal inputs to $u_{1941'}$ it is true that $u_{1941'}$ depends on both, but neither alone suffices. But then how t_{1941} is at H_{1941} is not sufficient for how t_{1941} is at H_{2062}, but how it is at H_{2062} still depends on how it is at H_{1941}. Why, then, is t_{1941} the same time at both H_{2062} and H_{1941}? Because there are no other hypertemporal causal relations between any other hypertimes and t_{1941}.

I grant that further work may need to be done to articulate the details on how objects are related through hypertime. It may turn out that some articulations are more palatable than others. For example, it may be easier to grasp how the temporal growing block endures through hypertime and gets truncated or expanded more rapidly via time travel than it is to see how pushing the button on the time machine at t_{2041}, H_{2041} causes t_{1921} (which last existed at H_{1921}) to become hypertemporally extended and exist again at H_{2042}. In other words, my model looks to have causation across both time and hypertime gaps.[2] Regardless of palatability, the issue for any such hypertemporal model is whether it is logically possible and models changing the past. Smith's arguments do not challenge the claim that they do.

2.2. Fixing the Past

Once we allow the possibility of changing the past does anything go? No, since we still cannot make, say, $u_{1981'}$ itself be contradictory. The best we can do is make 1981 one way at one hypertime and another way at another hypertime. But what about the

following popular time travel plot lines. Despite how careful the time travelers were, they did something that changed the past such that when they arrive back in the future things are drastically (and problematically) different. So back they go to fix their mistake. Or the time travelers deliberately set out to change the past—say kill Hitler as an infant to prevent the Holocaust, but when they arrive back in the future, they find an even worse history awaiting them. Oops! Back they go again. Or the bad guys temporarily get the upper hand by changing some key event in the past, so the good guys set out to fix it. The common want in all of these scenarios is the desire to undo the initial change and to fix the past. Is it possible to fix the past?

In [14], I implied a 'no' answer. But now I want to be more careful. The answer depends on what counts as an acceptable fix. For example, if the paint on my house is old and peeling, I might scrape the peeling paint down to the siding and try out a new color paint. My wife looks at the new color on the house and decides it is not nearly as good at it looked on the little color cards. She says she wants the old color back. My solution—put on a new coat of paint in the original color. Similarly, I can fix the dishwasher by taking out the defective part and putting in a working replacement part.

Time travelers, in hypertemporal models, can accomplish these sorts of fixes on the past as well. I can go back and kill grandfather before my mother was born, jump back to the present and find things horribly wrong. Back I go and intercept myself before the fatal shooting of grandfather. Have I made things *exactly* like they were before any time travel occurred? No. The first version of events contained no time traveler arriving to kill grandfather, but the successful fix contains two time-travelers—one potential killer and one preventer. But I haven't made my house exactly like it was before either, even if the new paint is the same color as the old paint—it is still new paint.

But could I get things back *exactly* the way things were before? There are two things one might mean by "exactly the same as before"—one could mean that the events of the most current hypertemporal chain of events are in one-to-one correspondence with the events of the original chain. On the other, one could mean that the universe just reverts to containing merely the original chain of events.

The first I suspect is logically possible—it is still a kind of change, but it might take God to help pull it off rather than just the efforts of any time traveler. The second, however, appears to be stronger than change. The time traveler does not want to change the first change—the time traveler wants the first change not to have happened at all–they want to not just change the past—they want to annul it. Can we give time travelers that?

3. Annulling the Past

To change the past is make the past different than it once was. For example, on hypertemporal models the time traveler might succeed in changing the past so that the past no longer includes the Holocaust. The time traveler might be motivated by the desire that her mother not have suffered so much during those years. But on a little reflection, our potential time traveler might conclude that changing the past is not enough. She does not merely want to change the past so that (hypertemporally) now her mother did not suffer then—she wants to make it such that no part of the past *ever* contains her mother suffering for that is what the time traveler wants to eliminate from the universe—the suffering.

To annul an event of the past is to make that event never have been a part of the past. Hanley [10], (p. 337) claims most time travel stories depict annulling the past, and also claims that annulling the past is impossible. Of course, he defines annulling the past as "making it the case that (unrestrictedly) some event both occurred and never occurred." But I take it our hypothetical Holocaust annuller does not, at least explicitly, have that contradictory want—she do not explicitly want the suffering to have both occurred and not occurred—she merely wants it to have never occurred.

Is annulling the past possible? Is it the case that we can make the past never contain the Holocaust? It depends on what we mean by 'never'. We could mean 'never' in the sense that the Holocaust is completely removed from the universe. If the Holocaust annuller is

interesting is eliminating the suffering, she can certainly want to have it just not be a part of the universe at all. Define strong annulment then as follows:

> *Strongly annulling the past:* To strongly annul the past is to make some event of the past never be part of the universe in its entire spatio-temporal expanse.

The universe is everything, not just everything now. Even if one is a presentist one can talk about the universe in its entire spatio-temporal expanse—it is just everything that was, is, and will be. Hypertimes are a type of time and so will be part of the temporal expanse of the universe. The universe is every single u at every single H at every single t regardless of which parts are 'real' or 'exist' or are 'accessible' at particular times or hypertimes. Strongly annulling the past then is to make some part of the universe in its entire spatio-temporal expanse not be part of the universe in its entire spatio-temporal expanse. As Hanley would put it, to make something that is unrestrictedly part of the universe, unrestrictedly not part of the universe.

Strongly annulling the past is logically impossible, since nothing can be unrestrictedly part of the universe and also unrestrictedly not part of the universe. We certainly cannot strongly annul the past on hypertemporal models. On such models the past becomes hypertemporally extended, but we do not remove or eliminate any of the hypertemporally past versions of the past. Think again of the model in terms of layers of paint. The original version of the past is the bottom layer of paint. Time travel to the past starts a new layer of paint over some portion of the bottom layer.[3] Traveling back again is just putting yet another layer over some portion of the previous layer. Hence, we can ask coherently whether a third Htu layer is in one-to-one correspondence event-wise with the bottom layer (at least the parts that overlap), but we cannot make the second Htu layer not be part of the universe.

Though there may be some debate about whether most time travel stories involve changing the past or as Hanley claims what I am calling strongly annulling the past, we should not conflate the two. Changing the past involves the past being one way and then another, annulling the past involves the past being one way and (unrestrictedly) never that way. The former is logically possible, while the latter is not. Yet, Sam Baron in "Back to the Unchanging Past" seems to make exactly this conflation. He writes: "A time traveler, Tim, **changes** the past when he brings it about that some event, object or property which is part of the past when he begins his journey through time is no longer a part of the past at the end of the his journey through time." [11] (p. 130, emphasis in original) So far so good. But he immediately continues:

> "Changing the past, as I will understand it, means 'overwriting' the past. A time traveler overwrites the past when they bring it about that an event E that (unrestrictedly) occurred before their journey through time (unrestrictedly) never occurred by the end of their journey through time."

In a footnote to the just quoted text, Baron writes: "I believe that changing the past in my sense is what Hanley (2009, p. 337) calls *annulling* the past . . . " Certainly, if we understand Baron's 'unrestrictedly' as 'part of the universe in its entire spatio-temporal expanse', then 'overwriting' and 'strongly annulling', and Hanley's 'annulling' are effectively the same.

But despite his claim that by 'change' he means 'overwriting', the definition he gives of 'change' is not the same as the definition he gives of 'overwriting'—to say that something is *no longer* one way is certainly different than saying it was *never* that way. It is true that the Blackburn Rovers are *no longer* a contender for winning the English Premier League (after all they play in the Championship League now), but it is false that they *never* were a contender for winning the English Premier League (since they won it during the 94/95 season). Hanley certainly does not conflate changing the past and annulling the past—for Hanley they are two separate categories. He thinks changing the past is possible. He also thinks that annulling the past more accurately captures what is going on in most time travel

stories. I disagree, but regardless, we both accept that changing the past and annulling the past are different things.

Indeed, if all Baron means by 'changing the past' is 'strongly annulling the past', then I agree with Baron that such 'change' is logically impossible and I certainly did not need any of Baron's arguments to convince me that that is impossible. I can think of no philosopher who presents a model purporting to model changing the past who was trying to model strong annulment.

There is, however, a weaker sense of annulment that is available as follows:

> *Weakly annulling the past*: To weakly annul the past is to make some event of the past never be part of the past.

Hypertemporal models can make a reading of that sentence come out true (but notice that the 'unrestricted' is gone and the 'never' will need to be read in a particular way.) Right now (t_{2021}) it is part of the past that Claus von Stuaffenburg is executed by the Nazis. Suppose instead of killing Hitler our time traveler rescues von Stauffenburg just prior to his execution and ferrets him to safety. After the time travel occurs and assuming no further time travel, in t_{2021} it is never (temporally) the case that Claus von Stuaffenburg is executed by the Nazis. After the time travel and successful rescue occurs there is no accessible momentary temporal slice that contains such an execution. On hypertemporal models what counts as the past (at a particular hypertime) are just those temporal slices that are both prior to the (hypertemporally) current temporal slice and accessible from the current slice. In other words, at a given hypertime, the past is whatever universe states are in the topmost Htu layer prior to that hypertime. So, referring back to Figure 2, at H_{2021}, the past includes u_{1901}- u_{2020}, but at H_{2142}, the past includes u_{1901}-u_{1920} and then $u_{1921'}$-$u_{2020'}$. As long as von Stauffenburg is never executed by the Nazis within the string of universe states accessible at H_{2142}, then it is true that the time traveller has made the (hypertemporally current) past such that von Stauffenberg is never executed in it. Hypertemporally of course it once (say t_{2021}, H_{2021}) was the case that von Stuaffenburg was executed by the Nazis, but now (say t_{2021}, H_{2142}) he no longer was executed by the Nazis. In other words, weak annulment is possible if 'never' is read temporally, but not if 'never' is read hypertemporally.

Baron's two definitions of 'change' and 'overwriting' might come out equivalent on hypertemporal models if we read them as follows: A time traveler changes the past if he or she makes it the case that some event that was (hypertemporally) part of the past is no longer (hypertemporally) part of the past. A time traveler overwrites the past if he or she makes it the case that some event that was (hypertemporally) part of the past now (hypertemporally) never (temporally) was part of the past.[4] But this weaker kind of annulment is just a kind of change and is certainly not capturing what Hanley intended by annulment. Recall we want to remove the suffering not just from the way the past is now, but from the universe entirely.

In response to a potential objection to one of his arguments against the possibility of strongly annulling the past, Baron writes:

> "At one hypertime, the past is one way—it features a war, say—and at a distinct hypertime, the past is a different way—it features no war. Add a time traveler who is responsible for this qualitative variation, and surely we can say that they've changed the past."
>
> "Again, I am willing to admit that this is 'change' in some sense. But it is not the notion of changing the past outlined in Section 2 [his definition of overwriting]. At best it is a version of the regular notion of change found in one-dimensional models of time." [11] (p. 141)

So far there is no problem, since that is exactly what changing the past advocates are trying to present—an account of the regular notion of change that successfully applies to the past. After all, we want to make sure we are still talking about change and not something else entirely. But after laying out the regular notion of change and showing

how it applies with hypertimes, Baron writes: "But all we have done is take the ordinary, unobjectionable kind of change—change that we always knew Tim could get up to in the past—and smear it out over a second, . . . dimension." [11] (p. 141)

Now we need to be extremely careful about what we are asking of Tim. Those who grant that time travel into the past is logically possible grant that Tim can affect the past, i.e., Tim can go into the past and interact with things. Tim can, for example, go back into the past and paint houses, turning a white house into a blue house. Tim can be the agent who changes a house in the past from being white to being blue. This is no different than what Tim can do in the present with no time travel at all. But we want Tim to be able to change the past—to go back in time to a house that was white throughout 1970 and paint it in June 1970 so that it is blue throughout the last half of 1970—that is what it would be to change the past and not merely affect it. And hypertemporal models provide a way to do exactly that—at one hypertime the past is such that the house is white all through 1970, but at a later hypertime, after the time travel, the house is white for the first half of 1970 but blue for the last half. Indeed, this is what I would describe as overwriting or undoing the past—the past was once one way and now we have overwritten it (or undone it) and made it another way. What we have not done (and cannot do) is make it such that the house was white throughout 1970 and now make it such that no part of the universe in any of its temporal, hypertemporal, or whatever extent is such that the house is white throughout 1970.

Strong annulment, with no restriction on 'never' is impossible—what was once part of the past cannot now never be part of the universe at all. Weak annulment is just a type of change and is possible according to the extant models of changing the past. We should not conflate the two.

4. Otherwising the Past

Strong annulment of the past is logically impossible. But then what are we to make of Peter Vranas' arguments about replacing the past? He writes:

> "Do I really want it to be the case that there is a first 1987 in which the declaration of love occurs and a second 1987 in which the declaration does not occur. No, I rather want it to be the case that the declaration never happened; I want it to be the case that there is a single 1987 in which, as a result of something I do in 2005, the declaration does not occur. To use a label, I want to replace the actual past." [12] (p. 371)

Terminology aside, what Vranas calls replacing the past certainly sounds like what I have been calling annulling the past. Vranas goes on to argue that (1) replacing is a kind of change and (2) that in fact it is the more interesting kind of change than the transforming kind of change I talked about in the previous section. Finally, he also argues that (3) if affecting the past is possible, then so is replacing the past. But since affecting the past seems to be the least problematic of all time traveler abilities, if Vranas is right, and if replacing the past is annulling the past, then it is also possible to annul the past!

I am not concerned with (1) since whether replacing is a kind of change is ultimately a terminological dispute about how to use the word 'change' and, with regards to actual usage, Vranas is right that we use 'change' in both the transformative and replacement senses. We can change light bulbs either by painting them or by unscrewing them and screwing in a new light bulb. Quibbling about what is 'really' change here is fruitless.

The only support that Vranas provides for (2) is an analogy with the future. He argues that talk of "changing the future is more interestingly understood as replacing than as transforming it. . . . Do I want to transform the future? No. I realize that such a desire would be incoherent (i.e., it could not possibly be satisfied.) I want instead to replace the future, to bring about a non-actual future, a future in which I don't die under torture." Similarly for the past: "Changing the past is more interestingly understood as replacing than as transforming it. I don't want to transform the past: I realize such a desire would be

incoherent. I want instead to bring about a non-actual past, a past in which I am born by Caesarean section." [12] (p. 374)

One might doubt that we really want to replace the future, but even if we do, does the analogy hold? One might say that it makes sense to desire replacing the future because the future is open, but does not make sense in the case of the past because the past is closed. What time travel allows, one might argue, is that I can hope to get to that past and make it other than it was before I pushed the button on my time machine—i.e., I can hope to transform it.[5]

Even if one insists that the analogy holds, I suspect there is a deeper problem with (2). To see the problem we need to examine Vranas' argument for (3)—If we can affect the past, we can replace it. Vranas argues for (3) from the single premise that "if it is possible to have a given causal effect on the past, then it is also possible to have a different, incompatible causal effect on the past." [12], (p. 377) Given that the effects are incompatible, they do not happen in the same possible world, and so it will be true that there is a possible world in which, as a result of my time travel, that world's past is different than the past of the actual world. For example, is it possible for me to kill Hitler before he becomes the leader of the National Socialist Party? Yes, even ignoring hypertimes, since there is a possible world in which I travel back in time and kill Hitler. Of course, that world is not the actual world, but rather a world in which Hitler never became the leader of the National Socialist Party. It is also quite likely a world in which I go back and kill Hitler for some reason other than his becoming the leader of the National Socialist Party and engendering the Holocaust, because those things do not happen in that world either.[6]

Surely that is too easy—is that really all it takes for it to be possible that I travel back in time and kill Hitler and prevent the Holocaust? Well, in a sense yes, but is it the sense Vranas said he was going to provide? No.

I grant that I have the ability to do something, kill Hitler before 1940 say, such that were I to do it, the past would have been different than it actually is. Call this ability, the ability to otherwise the past. (A similar ability holds for the future.) Whether I can otherwise some event of the past depends on whether there is some possible world in which, as a result of something I do in that possible world, the past there is other than it is in the actual world.

Is otherwising the past the same as, Vranas' terminology choice aside, replacing the past? No. To replace a light bulb is to remove the old bulb and put another one in its place. To replace a government is, in one form at least, to remove the elected officials from office and put other officials in their place. If we are really talking about replacement, which is a kind of change, then to replace the past, or a part of the past, is to remove that part and insert a different part. But that is certainly not what happens in the world in which I kill Hitler. I do not remove some part of the past and insert a new version.[7] I kill Hitler and there is no Holocaust—but there never was a Holocaust in that universe. (Maybe for good measure I get Stalin in that universe too.)

I grant that if I can affect the past, then I can otherwise the past. But otherwising the past is not replacing the past. Nor is otherwising the past annulling it, since I do not make some event that was part of the past (in the actual world) not be a part of the (actual) universe at all. If we consider the other possible world in which I do kill Hitler, I do not make some event that was part of the past in that world to not be part of the past in that world—I just make the past of that world as it always was. But then Vranas has not shown that what we would normally think of as replacing the past, and so annulling the past, is possible.

Is otherwising the past even interesting? Is it really, as Vranas suggests, under the guise of 'replacement' at least, what time travelers want? No. Imagine that you want to remove all the suffering of your mother during the Holocaust. Lucky you, I happen to have here a time machine that allows unrestricted (temporal) time travel. Even better, I tell you that you have the ability to otherwise the past (though to make the ability more palatable I may talk about being able to replace the past). You have the ability to do something, such

that if you were to do it, the past would be other than it actually is. You doubt that you have such an ability, so to reassure you, I provide Vranas' arguments and examples. You admit you have that ability[8], but are crestfallen. Why? Because you realize that, even with the time machine, and the ability to otherwise the past, you cannot make *this* universe such that your mother's suffering is not a part of it.

5. Conclusions

Sixty years ago, the philosophical consensus was that time travel to the past was logically impossible. Carefully distinguishing changing the past from affecting the past, however, shifted the consensus toward the logical possibility of time travel to the past. Even more recently some philosophers have argued that changing the past is logically possible. Whether that becomes the new consensus position on time travel remains to be seen, but at the very least it, if what I have argued here is correct, then there are models that model changing the past without being avoidances of the past. At the same time there are still limits on what is possible, even if we can, via time travel, at least in certain sorts of universes, change the past. We cannot, as Hanley notes some time travelers might genuinely want, strongly annul the past or fix the past by strongly annulling some unfortunate change already made.

Even Vranas' 'replacing the past' which I have labelled 'otherwising the past' does not provided us a way to give Hanley's most demanding time travelers what they want, for otherwising is not the sort of change we typically call replacing nor is it strong annulment. In fact, I see no way to give these time travelers what they want—to make it such that some part of the past is no longer part of the universe, in all its spatial, temporal, hypertemporal, or whatever extent, at all. To be subject to strong annulment, the part of the past must be a part of the universe. Being part of the universe in all its extent, it cannot also not be a part of the universe. Strongly annulling the past is logically impossible. To make one's greatest foe or the Holocaust to have never been a part of the universe is impossible. Even with time travel, we are too late.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 Task Two: "we need to give substantive content to the claim that some normal time t1 at hypertime a and some normal time t2 at hypertime b are or are not (hypertemporal parts of) the same normal time." [9] (p. 684).

2 Though once you are fine with causation across temporal gaps because of time travel itself, I am not at all sure what the extra burden is for causation across a hypertemporal gap.

3 In [14] changes made in the past propogate everywhere hypertemporally instanteously. Hence, time travel to the past terminates the lower layer. But if you change the rate of propogation you can get models that allow different things. For example, if the changes propogate forward one temporal instant per hypertemporal instant, then the original timeline would keep progressing forward. In effect you would have a growing universe with two moving presents. The bottom layer would be progressing forward at one time per hypertime, and a hundred years back on the timeline, the second layer would also be progressing forward at one time unit per hypertime overwriting the bottom layer as it progressed.

4 For the two definitions to be equivalent we might also need a pretty stringent view of event individuation. Even if the time traveller ferrets von Stauffenburg to safety, if ex-Nazis hunt him down and execute him in 1950 say, we might not be able to say that the execution of von Stauffenburg by the Nazis is never part of the new past. But this ultimately hinges on how we are individuating events. The execution as it occurred originally has been eliminated from the timeline, but is the new execution a new version of *that* event or a completely different event? However this issue is ultimately resolved, 'overwriting' would at worst still be a subclass of 'change'.

5 On hypertemporal models it actually does make sense to desire a transformation of both the past and the future.

6 I do not say impossible, since God could give me, in the dead Hitler world, a vision of how things are in our world, such that that vision motivates me to go back in time and kill Hitler.

7 Just as there is a weak sense of annulling that comes out true on hypertemporal models there is a sense of replacement that is possible on hypertemporal models. From the hypertermporal perspective it looks like we transform the past—we make u_{1941} into $u_{1941'}$. But temporally you might say that u_{1941} gets replaced by $u_{1941'}$.

[8] I also grant that another consequence of Vranas' arguments might be that one doubts the efficacy of possible world models of 'ability' or that it is capturing the time traveller's relevant abilitities. I take no stand on that possibility.

References

1. Gernsback, H. The Question of Time Traveling. *Sci. Wonder Stories* **1929**, *7*, 610.
2. Nahin, P. *Time Machines*, 2nd ed.; Springer: New York, NY, USA, 1999.
3. Hospers, J. *An Introduction to Philosophical Analysis*, 2nd ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 1967.
4. Bradbury, R. A Sound of Thunder. *Planet Stories* **1954**, *6*, 4–11.
5. Dwyer, L. Time Travel and Changing the Past. *Philos. Stud.* **1975**, *27*, 341–350. [CrossRef]
6. Norden, E. The Primal Solution. *Mag. Fantasy Sci. Fict.* **1977**, *53*, 133–156.
7. Locke, R.D. Demotion. *Astounding Sci. Fict.* **1952**, *50*, 71–80.
8. De Camp, L.S. Aristotle and the Gun. *Astounding Sci. Fict.* **1958**, *60*, 67–98.
9. Smith, N.J.J. Why Time Travellers (Still) Cannot Change the Past. *Rev. Portueguesa Filosophia* **2015**, *71*, 677–694. [CrossRef]
10. Hanley, R. Miracles and Wonders: Science Fiction as Epistemology. In *Science Fiction and Philosophy*; Schneider, S., Ed.; Wiley-Blackwell: West Sussex, UK, 2009; pp. 335–342.
11. Baron, S. Back to the Unchanging Past. *Pac. Philos. Q.* **2017**, *98*, 129–147. [CrossRef]
12. Vranas, P.M. Do Cry Over Spilt Milk: Possibily You Can Change the Past. *Monist* **2005**, *88*, 370–387. [CrossRef]
13. Meiland, J. A Two Dimensional Passage Model of Time for Time Travel. *Philos. Stud.* **1974**, *26*, 153–173. [CrossRef]
14. Goddu, G.C. Time Travel and Changing the Past. *Ratio* **2003**, *16*, 16–32. [CrossRef]
15. Van Inwagen, P. Changing the Past. In *Oxford Studies in Metaphysics*; Zimmerman, D.W., Ed.; Oxford University Press: Oxford, UK, 2010; Volume 5, pp. 3–28.
16. Hudson, H.; Wasserman, R. Van Inwagen on time travel and changing the past. In *Oxford Studies in Metaphysics*; Zimmerman, D.W., Ed.; Oxford University Press: Oxford, UK, 2010; Volume 5, pp. 41–49.
17. Bernstein, S. Time Travel and the Movable Present. In *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen*; Keller, J., Ed.; Oxford University Press: Oxford, UK, 2017; pp. 80–94.
18. Effingham, N. The Metaphysical Possibility of Time Travel Fictions. *Erkenntnis* **2021**. [CrossRef]
19. Goff, P. Could the Daleks Stop the Pyramids Being Built? In *Dr. Who and Philosophy: Bigger on the Inside*; Lewis, C., Smithka, P., Eds.; Open Court Press: Chicago, IL, USA, 2010; pp. 67–74.
20. Wasserman, R. *Paradoxes of Time Travel*; Oxford University Press: Oxford, UK, 2018.
21. Effingham, N. Vacillating time: A metaphysics for time travel and Geachianism. *Synthese* **2021**. [CrossRef]
22. Goddu, G.C. Avoiding or Changing the Past. *Pac. Philos. Q.* **2011**, *92*, 11–17. [CrossRef]
23. Smith, N.J.J. Bananas Enough for Time Travel? *Br. J. Philos. Sci.* **1997**, *48*, 363–389. [CrossRef]

philosophies

MDPI

Article

Self-Fulfilling Prophecies

Stephanie Rennick

Department of Philosophy, University of Glasgow, Glasgow G12 8QQ, UK; stephanie.rennick@glasgow.ac.uk

Abstract: Causal loops are a recurring feature in the philosophy of time travel, where it is generally agreed that they are logically possible but may come with a theoretical cost. This paper introduces an unfamiliar set of causal loop cases involving knowledge or beliefs about the future: self-fulfilling prophecy loops (SFP loops). I show how and when such loops arise and consider their relationship to more familiar causal loops.

Keywords: causal loops; self-fulfilling prophecies; time travel; foreknowledge; coincidence; inexplicability

Citation: Rennick, S. Self-Fulfilling Prophecies. *Philosophies* **2021**, *6*, 78. https://doi.org/10.3390/philosophies6030078

Academic Editor: Alasdair Richmond

Received: 3 August 2021
Accepted: 9 September 2021
Published: 18 September 2021

> *Joakim stands in front of his wardrobe, indecisively. In a flash of inspiration, he selects the black kimono. After all, he ponders, he already knows that's what he'll wear: the Delphic Oracle texted him as much that morning. Saved the mental labour of choosing an outfit, his mind wanders: 'she knew, because I'm wearing it; but I'm wearing it, because she knew that I would . . . '*

Causal loops are a recurring feature in the philosophy of time travel literature, where they are commonly divided into two varieties: object loops and information loops [1–3]. Here I discuss a different set of causal loop cases that involve knowledge or beliefs about the future: loops that arise in a subset of what I call self-fulfilling prophecies (SFPs)[1]. Although SFPs are a popular trope in the folk canon, they have yet to be the focus of detailed philosophical treatment. This paper serves as a first map of the conceptual terrain of SFPs, identifies their structure and features, and considers their relationship to more familiar causal loops.

I start by introducing causal loops more generally (§1) and then turn to introducing SFPs (§2). The rest of the paper deals with potential worries, including the apparent inexplicability (§3) and improbability (§4) of SFP loops. I will argue that neither worry is insurmountable, and that we should consider at least some SFPs even less problematic than the causal loops we are more accustomed to.

1. Causal Loops

A causal loop is a chain of events where each "is one of the causes of the next event, and whose last event . . . is one of the causes of the first event" [2] (p. 259). Events in a loop need not be complete causes or effects of one another, but may be:

> In a causal loop, the arrows of causation go around in a circle, but there might be additional arrows that lead into the circle, or arrows that lead out of it. If there are no such branches then the loop is said to be *causally isolated* [2] (p. 259).

Causal loops crop up frequently in the time travel literature, most famously described by Lewis as,

> Closed causal chains in which some of the causal links are normal in direction and others are reversed . . . Each event on the loop has a causal explanation, being caused by events elsewhere on the loop. That is not to say that the loop as a whole is caused or explicable. It may not be [1] (pp. 148–149).

A particularly memorable example comes from French and Brown: an archaeologist travels several millennia into the past in an attempt to discover the origins of a recently unearthed human skeleton, only to die and become that skeleton [4] (p. 208). This scenario involves a causal loop—discovering the skeleton results in the archaeologist's time travel, which results in his death, decomposition, and eventual discovery by the archaeologist. An important element of causal loops is the (weak) predestination they entail: the fact that the archaeologist discovers the skeleton-that-is-himself seems to entail that he will travel in time (and die, and decompose etc.). It is true[2] that the archaeologist, upon finding the skeleton, will go on to travel in time. Note that this isn't a fatalistic conclusion—the archaeologist did not *have* to time travel, but the fact he does results in the skeleton's existence (and importantly, identity: if he was not going to travel in time, he still might have found a skeleton, just one belonging to somebody else.)[3]

While some have disputed the possibility of causal loops depending on their accounts of time and causation[4], most philosophers of time travel think them at most weird [2,3,13,14]. As Lewis remarks,

> Strange! But not impossible, and not too different from inexplicabilities we are already inured to [1] (p. 149).

However, time travel is not the only context in which causal loops arise. Meyer notes a second:

> The other cases involve models of the general theory of relativity—first discussed by Kurt Gödel (1949)—that possess closed timelike curves in which time itself loops along a particular wordline. In such models, there is no backwards causation and travelling back in time requires no particular effort; one just has to follow an appropriately chosen worldline [2] (p. 259).

Skrzypek adds a third:

> [Loops] can be constructed by granting the existence of some causal agent existing in eternity, something or someone that has equal and simultaneous access to events at several times. In such cases, either the causal efficacy of the later event could "run through" the causal efficacy of the eternal being to the earlier event, or the causal efficacy of the eternal being could "run through" the causal efficacy of the later event and back through eternity to the earlier event [15].

The example he focusses on is the (later) impact of Jesus's life, passion and resurrection on the (earlier) immaculate conception of Mary.

The loops this paper focuses on cross-cut those identified by Meyer and Skrzypek: although many of the cases I discuss are time travel stories involving backwards causation, not all need be: foreknowledge-generating loops could also arise via backwards causation without time travel, or with forwards-causation via closed time-like curves (CTCs). Some philosophers of religion have worried that causal loops might result from a particular kind of divine foreknowledge called *simple foreknowledge* [16–18]. Under this account, God's witnessing the future is akin to us peering out a window: he has immediate, direct access to truths about the future[5]. Suppose that God today knows one of his faithful followers (FF) will die on Wednesday, so decides to warn him on Monday that he should tidy up his affairs. As a result of the warning, the follower gets very panicky, culminating in a heart attack on Wednesday (see Figure 1).

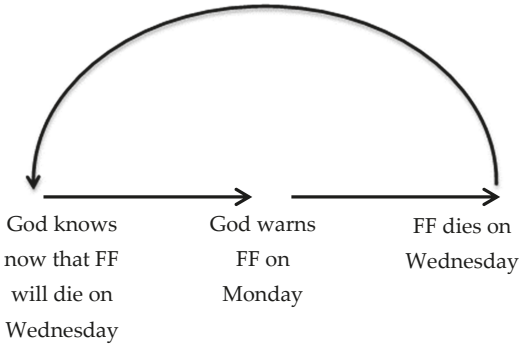


Figure 1. Divine foreknowledge in a causal loop (arrows denote causation).

Interestingly, considerable philosophical labour has been expended trying to prevent a connection between divine foreknowledge and causal loops, by—for instance—limiting God's knowledge to certain facts or certain times ('bracketing' God's knowledge) or making it the case that God only has knowledge after certain choices have been made. The worry centres around what Hunt calls 'the metaphysical principle' (MP):

MP: It is impossible that a decision depend on a belief which depends on a future event which depends on the original decision [16] (p. 486)[6].

MP-violating scenarios are not only causal loops, they are SFPs, and so they are loops of the particular variety we are interested in here. But as will become clear, you do not need God to get them.

2. Self-Fulfilling Prophecies (SFPs)

SFPs occur when knowledge or awareness of the future—perhaps by prediction, testimony, revelation or observation—is a crucial factor in future events occurring as the 'prophecy' describes. In a literal example of 'prophecy', Oedipus kills his father and marries his mother as a result of actions to avoid fulfilling a prophecy in which it is foretold that he will kill his father and marry his mother [19][7]. He knows the content of the prophecy and by trying to thwart it inadvertently ensures its veracity. SFPs are particularly vivid examples of how foreknowledge can impact the causal chain of events the knowledge describes: had he not heard the prophecy, it seems unlikely that Oedipus would have behaved as he did. The conundrum resulting from the feedback of foreknowledge on present action is vividly depicted in the following interaction from *The Matrix:*

> The Oracle: [. . .] And don't worry about the vase.
>
> Neo: What vase?
>
> [Neo knocks a vase to the floor]
>
> The Oracle: That vase.
>
> Neo: I'm sorry.
>
> The Oracle: I said don't worry about it. I'll get one of my kids to fix it.
>
> Neo: How did you know?
>
> The Oracle: What's really going to bake your noodle later on is, would you still have broken it if I hadn't said anything [21].

(Note that SFPs need not involve agents actively trying to *thwart* the events in question, although in examples from popular culture they often do; more on this below.)

Some SFPs are causal loops, and some are not. Suppose I say to you, "I know you will wear purple tomorrow", and then you wear purple because I said you would. By the definition given above, this is an SFP. If, however, I have just pulled 'purple' from the aether, then even if it turns out to be true that you wear purple, there is not a loop here[8]. There are

lots of examples like this in "The Fortuneteller": the town's fortune teller, Aunt Wu, makes a series of predictions that turn out to be correct, although the villagers fail to recognise that her accuracy is the result of their actions, rather than her having actual foreknowledge. For instance, Wu tells an old man that on the day he meets his true love he will be wearing red shoes, and as a result, he dons red shoes every day [22]. This is a standard causal chain: the man's actions result from what he perceives to be foreknowledge, but they do not influence Wu's prophecy (Figure 2):

Figure 2. A Simple Causal Chain.

(Note that had he not believed the prophecy—had he not given it sufficient credence—it would not have been self-fulfilling; more on this below).

We can also have non-loopy scenarios with the causal chain running in the opposite direction if agents fail to act on the basis of their foreknowledge (these aren't self-fulfilling). Recall the case of the faithful follower (FF) from §1, where God knew the FF would die, warned him in advance, and thereby caused his heart attack and ultimate demise. Now suppose that God decides not to intervene: the death of FF causes God to know that FF dies, but the causal connection is one-way (as shown in Figure 3):

Figure 3. Divine foreknowledge without loop.

Sometimes, though, SFPs form causal loops. For instance, in Garth Nix's *Old Kingdom* Trilogy, the prophetic Clayr, in the present, have visions of inducting new members; then in the future, they induct those members based on the previous vision [23] (p. 15f). But they only have the vision because the new members will be inducted (as depicted in Figure 4):

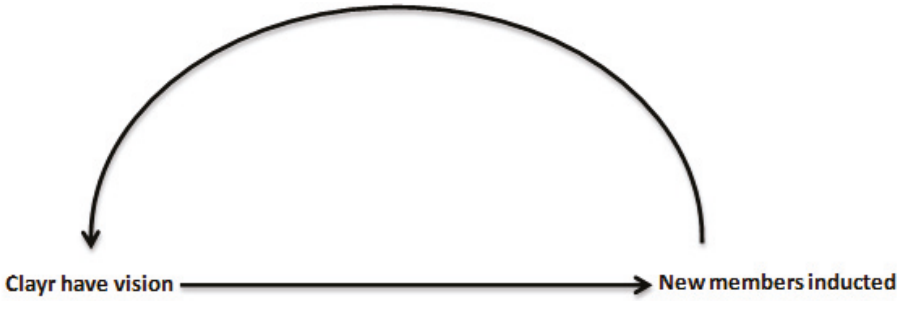


Figure 4. A Causal Loop[9].

Despite the familiarity of these scenarios from popular culture, mythology and religion, foreknowledge-generating causal loops (i.e., SFP loops) have not received detailed treatment in the philosophical literature. This might be surprising; indeed, three names might come to mind as potential counterexamples—Lewis, Goldman and Hanley. However, while Lewis briefly discusses both loops and fatalism at different points in "The Paradoxes of Time Travel" [1], that foreknowledge might result in such loops is never suggested[10]. Goldman on the other hand does provide an excellent example of a foreknowledge-generating loop, as described below, but his agenda is entirely different: to rebut what he calls 'anti-predictionism', i.e., an objection to determinism based on the impossibility of predicting voluntary actions[11]. Hanley comes the closest: he discusses loops involving intentional ac-

tion and the coincidences they might involve, and I will refer to his arguments throughout. However, it is worth noting that the role of belief in his cases goes unmentioned.

Goldman's Book of Life thought experiment goes as follows:

> While browsing through the library one day, I noticed an old dusty tome, quite large, entitled "Alvin I. Goldman". I take it from the shelf and start reading. In great detail, it describes my life as a little boy. It always gibes with my memory and sometimes even revives my memory of forgotten events. I realise that this purports to be a book of my life . . . I look at the clock and see that it is 3:03 . . . I turn now to the entry for 3:03. It reads: *"he is reading me. He is reading me. He is reading me* [26] (p. 144).

Like the Clayr case, this scenario involves a causal loop: at t_0, Goldman reads a page in the book of life that describes events taking place at t_1. At t_1, after the reading, Goldman performs an action φ. Suppose that the author of the book gains knowledge of Goldman's φ–ing (there are various ways to conceive of this, from time traveller to fortune teller) as a result of Goldman's φ–ing. That is, the book records that Goldman φs because he φs. But Goldman may, at least in part, φ as a result of this knowledge (or in defiance of the knowledge if it is a thwarting case[12]), and thus the knowledge is a contributing factor in the causal chain of events. To make this clearer, consider the following:

> I now turn to the entry for 3:28. It reads, "He is leaving the library, on his way to the President's office." Good heavens, I say to myself, I had completely forgotten about my appointment with the President of the University at 3:30 . . . Since I do have a few minutes, however, I turn back to the entry for 3:22. Sure enough, it says that my reading the 3:28 entry has reminded me about the appointment [26] (p. 144).

Why did he leave the library at that time? Because it was written he would. Why was it written? Because he left at that time. Like the archaeologist case, there is (weak) predestination at play here: the fact that Goldman reads a book which truthfully describes his future actions entails that he performs such actions. It is true at 3:22 pm that Goldman will leave the library at 3:28 pm. He did not have to, but the fact that he does makes the book true (and report that detail, instead of a variation).

However, not all of the events mentioned in the Book of Life case are loopy: neither those that occurred before Goldman started reading, for instance, nor any that might have occurred due to Goldman's misreading (or misremembering) of the text. In those cases, Goldman's actions are either causally unrelated to the prediction, or themselves the cause of the predictions (via backwards causation, a CTC or similar). And the claim above that 'Goldman's action makes the book report that detail' is predicated on the assumption that the writing in the book came about as a result of Goldman's actions: as if a fortune teller, God, or time traveller witnessed Goldman performing said actions and recorded them accordingly. But if this is not the case, then the book causes Goldman to remember his appointment, but his remembering his appointment has no bearing on the book's contents[13].

As Goldman's Book of Life and the Clayr examples demonstrate, in some cases knowledge of the future can influence actions in the present: the knowledge bears on the causal chain of events it describes. Sometimes (but, as demonstrated, not always) this leads to a causal loop[14].

However, as I have alluded to throughout, while many SFP scenarios do involve foreknowledge (at least under standard accounts of knowledge), we need not meet the bar of 'knowledge' for such a loop to arise [28] (p. 63). What SFP cases have in common is that awareness of (and belief in) the content of the prophecy leads to its coming true. In the story of Oedipus, for example, we are led to believe that the protagonist would never have killed his father and married his mother if he had not heard and believed the prophecy in the first place: it is the prophecy that provides the impetus, and serves as a catalyst, for the events that follow[15].

In most cases, at least one character in an SFP scenario (either the prophet or the person to whom the prophecy pertains[16] has a true belief about the future. For instance, the Clayr have a true belief that they will induct certain recruits into their ranks. But even this is not required; one could imagine a scenario where the predictor has access to and reports a truth about the future without believing it, where the subject of the prediction has a false belief about the future, and an SFP (even a loopy SFP) still arises. For instance (I call this case "A Comedy of Cellars"):

> Julia peers into her crystal ball and witnesses a future so surprising that she can't believe it to be the case; nonetheless she's a diligent sort and records her vision in her diary. Her assistant, Sue, glances in the diary at the end of the day, but isn't wearing her glasses so misreads Julia's handwriting, forming the false belief that her twin sister Prue will end up trapped in the cellar. Sue avoids the cellar as she is deathly afraid of the dark, but ventures down to save her more adventurous sibling. In her haste she forgets the keys, thereby trapping herself in the cellar and proving true the vision that Sue—not Prue—would be stuck in a place Julia would never expect her to tread.

In this case and other SFP cases, the belief (albeit false here) plays a significant role in events transpiring as they do. Nonetheless, we might think that cases such as this one are so improbable as to occur only very rarely. I consider the coincidences involved in SFP loops in §4. First though, I address another worry. One might wonder, in this case and the others discussed, *why* the sequence of events comes about: whether we can adequately explain where the knowledge, belief or decision comes from. This is a common worry about certain causal loops, which I turn to now.

3. Inexplicable Loops

There are several (interrelated) ways in which the charge of inexplicability is levied against causal loops. For instance, one of the main objections in the literature on the simple foreknowledge case outlined in §1—where God knows his faithful follower will have a heart attack and, in telling him, brings about the events and thus his knowledge—is that such loops serve as a bad explanation for why events occur [18]. In the time travel literature, it has been observed that while each event in a loop has a causal explanation, as it is caused by other events in the series, the loop as a whole may have no cause and thus be inexplicable. Finally, the content of at least some causal loops is argued to arise *ex nihilo*, and this is also claimed to be inexplicable. I find myself unsure as to whether SFP loops fall into this latter category, so instead will here make a disjunctive claim: if they do not, then SFP loops are even less troublesome than some other kinds of causal loops. However, even if they do, the time travel literature gives us good reason to resist that inexplicability worry, along with those levied against causal loops more generally.

An important feature of some, but not all, causal loops is the lack of origin of their content. Take, for instance, the main character in Robert Heinlein's *"-All You Zombies–"*, who is—thanks to time travel and a mid-life sex change—his own mother, father, daughter and son [31][17]. Their genetic information is caught up in a closed causal loop, with no apparent origin: it comes 'from nowhere'. Not just information can be loopy in this way, objects can too; imagine a time traveller who goes back in time and gives his younger self plans to build a time machine. The younger self grows up, builds the time machine, and goes back to give himself the plans[18]. As with Heinlein's character, the quandary lies in the blueprints' origin: where did they come from in the first place? So as not to confuse these with other types of causal loop, I call them CEN loops: loops that involve information or objects created (or appearing) *ex nihilo*[19]. In time travel scenarios, all sorts of objects, information, and even people, can have a loopy causal origin[20].

It is generally agreed in the literature that CEN loops are at least logically possible, although for other reasons they might be unpalatable; as Meyer notes, "loops are widely thought to constitute a theoretical cost of any view that permits them" [2] (p. 260).

At least some SFPs look to be CEN loops: the content of the Clayr's knowledge might seem to come 'from nowhere', and likewise for Goldman's decision to leave the library. Here's another example to consider:

> Billy is a contestant on a game show with very similar mechanics to the Newcomb Problem: he has a choice between one box and two boxes, and a highly-accurate predictor will predict his choice in advance. However, this particular predictor has access to the future (by time travel, crystal ball or a Gödelian telescope), and this is why she is so accurate—she witnesses the future choice, and thereby knows what Billy will choose. Billy is a stalwart two-boxer: in every conversation with friends and colleagues prior to the game he has insisted he will pick both boxes; he dreams at night of picking both boxes etc. However, the tables are turned when the predictor reveals her prediction to him prior to his choice: she says he will pick one box. As a result of this revelation, Billy decides to pick one box—after all, he reasons to himself, he now knows he will (see Figure 5)[21].

Where does the content of the knowledge come from? Well, the predictor knows that Billy will pick one box, because Billy picked one box. But Billy picked one box, because the predictor knew he would pick one box. If you do not like the word 'know' here, replace with 'believe': the predictor believed Billy would pick one box, as she witnessed Billy pick one box. But Billy picked one box, because she believed he would (or more long-windedly, Billy picked one box because he believed that she believed he would pick one box). Each link on the chain is explicable in terms of the previous link, but the loop as a whole, and the information contained in the foreknowledge (i.e., the outcome of Billy's decision), appears not to be.

In his "Paradoxes of Time Travel", Lewis concedes that CEN loops are inexplicable but remains unperturbed, describing their possibility as "remarkable" and arguing that information arising from nothing is no different than the many other inexplicable phenomena we manage to accept, such as the "decay of the tritium atom" [1] (p. 149). Nonetheless, he considers loops with information for content—like our SFPs might be—especially remarkable, asking

> Where did the information come from in the first place? Why did the whole affair happen?
>
> And concluding,
>
> There is simply no answer [1] (p. 149).

Hanley disagrees, arguing that the question, 'Where did it come from in the first place?' is malformed; it is unanswerable, but only because there is no first place to talk about on a loop. Instead, he suggests, "the well-formed question 'Where did the information come from?' has a straightforward answer: from itself, by completely ordinary causal means" [3] (p. 137).

Perhaps the best articulation of why the bizarreness of CEN loops is merely apparent is found in Levin, who argues that questions about the origin of objects or information caught in a loop are

> No different from questions about where anything originally came from. We can ask about the origin of the atoms that make up [the time traveller]; their timeline is not neatly presented to us. The atoms either go back endlessly, or if the universe is finite, they just start. In either case the question of ultimate origin is as unanswerable as the question of the [loop contents'] origin. What makes us think that when such questions are asked about the loop they are different and ought to be answerable is that the entire loop is open to inspection. Sub specie aeternitatis this difference disappears [37] (p. 70)[22].

That is, we do not expect to be able to explain the causal history of the atoms that make up the objects surrounding us as they stretch back so far in time (and perhaps endlessly).

By contrast, in Billy's case, we have the entire causal history of the information (or decision) open to us. Thus we expect to be able to determine how or why it came about[23].

In terms of explicability, we can say the same things about loopy SFPs as has been said in the time travel literature. Each event in the loop comes about (at least partly) as a result of its predecessor. Each event is as explicable as events in a linear causal chain, with the added benefit that there are no uncaused first causes. There may not be an explanation for why the loop as a whole came about, but that is not unique to causal loops: if it is a problem, it is one the universe, God, and—as Levin notes—atoms, face as well.

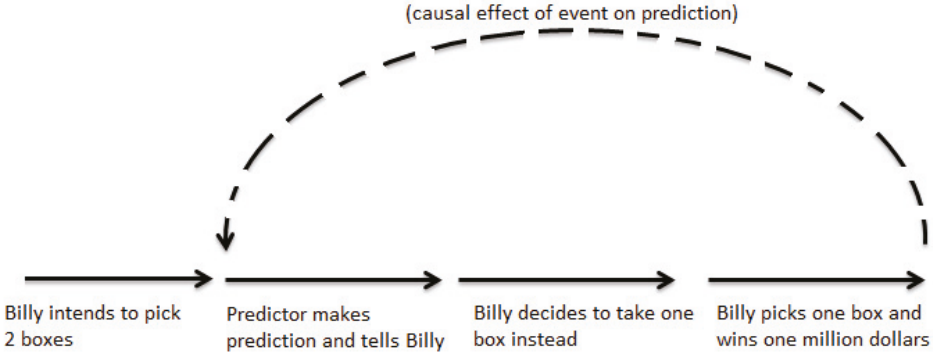


Figure 5. A candidate CEN loop.

4. Improbable Loops

> The really puzzling question is, 'Why does the information work?' . . . It's no mystery how the information works . . . The interesting question is why this is so. And the only possible answer is: coincidence [3] (p. 138).

The logical possibility of causal loops has been established decisively elsewhere, but one might still be concerned that cases like those I have introduced are so improbable that they nonetheless impose a significant theoretical cost. The kinds of coincidences involved in causal loops—and CEN loops in particular—have been discussed in various places in the time travel literature, and it is generally agreed that how improbable they are depends at least in part on the content of a given loop [3,14,24]. I will not replicate this discussion here, but I do want to say a little bit about why SFP loops, even if they involve content arising *ex nihilo*, might be less improbable or only require more ordinary coincidences than some other loops.

There are three reasons for this. Firstly, SFP loops are not object loops (but rather a subset of, or at least more akin to, information loops, to follow Lewis and Hanley's dichotomy). Objects age, they experience wear-and-tear, and if caught in a loop they must reverse age at some point so as to end up—atom-for-atom—as they started (Hanley calls this the 'restoration problem') [3] (p. 131f). SFPs are not subject to this problem, and thus do not require coincidences to surmount it. Secondly, the kind of *ex nihilo* content in the SFP cases presented above is in general simpler than that featured in the more extreme time travel loops; there are exceptions to this, and nothing relies on it, but it seems at least plausible that the kind of information we might come to know (e.g., that I will wear a red dress next Friday) is less complex than—for instance—the complete genetic information that makes up a person[24].

Finally, and most significantly, recall that one of the ways SFPs can be self-fulfilling is due to complying agents. Take the case of the Clayr (depicted in §2 Figure 4): the Clayr have a vision of the future in which they induct new members into their ranks. When the foretold day arrives, they induct these members. Given their visions are usually reliable, there is nothing coincidental about them seeing the new members being inducted. Likewise,

there is nothing coincidental about them inducting those members: they did so on the basis of the vision. At most, as Hanley notes of an analogous time travel case, "it is coincidental that they happen to know what they need to know, in order to do what they do", but "the sorts of coincidence involved are completely ordinary" [3] (p. 136).

Certainly there are SFP cases in fiction, or that we could imagine, that involve a great number of coincidences: those, for instance, where the agent hears a prophecy and wishes to foil it, only for circumstance to conspire such that they end up bringing about what they wanted to avoid[25]. "A Comedy of Cellars" in §2 might be one such example. Just as when we fail in more mundane tasks—like trying to get to work on time—some of the coincidences that scupper us will be ordinary (a run of red traffic lights, an unscheduled call from an overly talkative relative) and some may be more improbable (a monkey escaping from the local zoo and running off with our keys). Even if this type of case—with an agent trying to avert a true prophecy—does involve highly improbable coincidences, we have no reason to believe that it would be representative of SFP scenarios[26]. Indeed, if the coincidences required in foiling-but-self-fulfilling-anyway cases are especially unlikely or more numerous as opposed to the compliance cases, it is likely that the latter would comprise the bulk of SFP loops[27]. If the Clayr hadn't (a) willingly complied with the vision, then either (b) wacky hijinks would have ensued such that they complied inadvertently (as in so many fictional cases) or (c) the vision would have proven false. Neither (a) nor (c) requires any surprising or improbable coincidences, and of those only (a) results in a loop. So, even if (b)-type cases are improbable and thus unlikely to occur, this has little bearing on the probability of SFP loops more generally. Hanley observes that intentionality could reduce the improbability of causal loops: "the existence of agency may be the very thing that permits causal loops to obtain" [3] (p. 148). Cases where agents aim to bring about what they know will occur are prime examples of this.

5. Final Thoughts

Scenarios where agents have future-directed knowledge or beliefs that come to affect their present action are interesting and until now, underexplored, sources of causal loops. At worst, SFP loops are as inexplicable and improbable as the more familiar causal loops appearing in the time travel literature, but I suspect that in fact they require less to get off the ground.

I want to finish with one last question: do SFPs only arise in cases where agents know (or have beliefs about) their own futures? The short answer is no. What follows is the longer answer.

It is helpful to differentiate between what we might call 'first-person' and 'third-person' foreknowledge. (For ease I'll say 'foreknowledge', but throughout 'forebelief' can be substituted). By first-person foreknowledge I mean cases where someone knows their own future, rather than someone else's[28]. Most instances of third-person foreknowledge will not result in loops, as the foreknowledge will be merely the result of events occurring as they do, rather than a cause of the events—that is, the foreknowledge describes the events rather than (contributing to) bringing them about. For example, if a time traveller knows I will wear a red dress next Friday because she saw me wearing it in the future, then my wearing the dress causes her foreknowledge (in addition, presumably, to other causes). However, if I remain unaware of her knowledge, it seems unlikely that it would influence my choice of attire, thus there is no loop (see Figure 3 for an analogous case).

In all of the examples mentioned thus far, it is specifically the object of the prophecy coming to learn of it (or mistakenly thinking they have learned of it) and putting some stock in it that leads to its coming true. Given this, it is reasonable to wonder if SFPs are by their nature restricted to the first-person: if the foreknowledge must be possessed by the focus of the foretelling, rather than (or in addition to) a third party.

Although many SFPs do take this form—and indeed, they seem to be the most common and vivid in the folk canon—the foreknowledge need not *strictly* be first-person. For instance, in *Kung Fu Panda*, Grand Master Oogway has a vision in which the villain,

Tai Lung, escapes his prison. Tai Lung is not privy to the content of the prophecy. In an attempt to thwart the villain, Master Shifu sends a bird to the prison to increase security, thereby providing Tai Lung with the means of escape: a feather for a lock pick [39]. This is nonetheless an SFP as Shifu's foreknowledge plays a crucial role in bringing about the events that were foretold, even though the stated subject of the prophecy—Tai Lung—is unaware of this.

However, one might argue that all SFPs are implicitly first-person. Shifu is not aware of the role he plays in Tai Lung's escape because he does not have complete information. If a complete prophecy had existed describing the full set of events, then his actions would be contained therein.

In this paper I have focused on knowledge and belief, but it might well be that other future-oriented mental states can generate causal loops. Historically, the kinds of causal loop scenarios that have typically been discussed—especially those containing objects or information arising *ex nihilo*—have involved the kinds of things we do not ordinarily expect to come from nowhere: time machine blueprints, a working set of human DNA, and so on. But SFPs suggest a different direction: less unusual stuff—like beliefs—can generate causal loops[29]. Hanley points to intentionality as another source. It remains to be seen what else among our mental furniture might be suitably loopy.

Funding: Funding was provided by the Swiss National Science Foundation (grant no. 182847).

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 It may be that these are best understood as a subset of information loops, but I won't assume as much. Either way, they have some features that aren't shared by information loops more generally, such as the role that belief plays in their occurrence.

2 Or will become true, depending on your theory of time. For ease I will assume four-dimensionalism as per Lewis [1], but much of what I say could be adapted to other theories of time.

3 Whether there are additional free will limitations with this kind of case, or backwards time travel more generally, is a bigger question than I have scope for in this paper. See for example [5].

4 For example, [6] (pp. 123, 175–177). Cf. [7] (Chapter 12); [8] (Chapter 17); [9] (p. 4 fn. 1). Mellor's arguments (and similar) have been challenged in various places, including [10] (pp. 131–134); [11,12].

5 As opposed to conditional or counterfactual knowledge, for example.

6 It isn't logically impossible, but may well be theologically impossible (if, as has been suggested, it undermines the providential usefulness of divine foreknowledge). More interesting for my purposes is the claim that MP-violating circumstances are a bad explanation for why events occur as they do, cf. [18]. This is discussed further in §3.

7 For a particularly explicit example see the conversation between Harry and Dumbledore on Voldemort's actions in [20] (pp. 740–741)

8 Whether or not this would count as a case of fore*knowledge* as opposed to just forebelief (i.e., the belief that you will wear purple, formed as a result of my testimony) will depend on the connection between truth and belief that your epistemology requires. Either way, it isn't a loop, as the causal chain runs one-way from the belief today to your getting dressed tomorrow. If, however, I said you would wear purple because I'm a time traveller from the future where I saw you wearing purple, then it'd be a loop.

9 This is simplified: we might expect intermediate steps such as 'Clayr decide who to induct' between their vision and the induction.

10 Of course, time travel resulting in knowledge or beliefs about the future and the consequences of that in terms of free will, intentionality and resulting coincidences are discussed in several places, including [3,24,25]. However, that knowledge of (or beliefs about) future facts can generate causal loops (whether in tandem with or independently of time travel) was not the focus of these analyses.

11 That the thought experiment he provides happens to be such a loop is grist to my mill, but undermines neither the novelty of this paper's conceptual mapping nor its conclusions.

12 NB. These are directly analogous to bilking cases in the time travel literature (Cf. [25]).

13 Likewise, if the author of the book was a supercomputer that could calculate the future based on deterministic laws and a complete understanding of the present (i.e., could have foreknowledge without backwards causation), this would not result in a causal loop.

[14] To think that the relationship between foreknowledge and foreknown events always results in such a loop is what Craig calls "a misunderstanding in which the causal relation between an event or thing and its effect is conflated with the semantic relation between a true proposition and its corresponding state of affairs" [27] (p. 337).

[15] Oedipus does not believe the prophecy is infallible, however, but rather that it will prove true *unless he acts to prevent it*. Another classic SFP case can be found in W. Somerset Maugham, "An Appointment in Samarra" from *Sheppey* (1933), as cited in [29] (p. 57): a merchant in Baghdad encounters Death and, based on what he perceives to be a threatening gesture, flees to Samarra to avoid his fate. Death notes that it wasn't a threatening gesture but "only a start of surprise. I was astonished to see him in Baghdad, for I had an appointment with him tonight in Samarra." If the merchant hadn't believed Death was out to get him, he'd have had no reason to go to Samarra. A nice parody occurs in [30] (pp. 77–78): Rincewind the wizard runs into Death, who comments that they have an appointment elsewhere soon and asks if Rincewind would mind going there. Rincewind declines.

[16] These might be the same person, e.g., the Clayr.

[17] Similar cases can be found in [14] and *Futurama*. Jane grows up in an orphanage; as a teenager she is seduced by a young man, falls pregnant and gives birth to a baby. Jane suffers trauma to her reproductive organs during labour, but doctors discover she is intersex and she undergoes sexual reassignment surgery. Now identifying as a man, Jane is taken back in time by a Bartender, where he meets and impregnates a young woman called Jane. The Bartender then recruits the young man to serve in the Temporal Bureau. The Bartender—revealed to be an older timeslice of the main character Jane—takes the baby back in time to an orphanage; he returns to the Bureau to contemplate his caesarean scar and the creation and recruitment of himself.

[18] This is a version of an example in [1] (p. 149).

[19] They are rarely given a name, but are mostly (including by Lewis) bundled together under the generic name 'causal loops'. Occasionally they are called 'ontological loops', a specific type of 'closed causal loop', or the 'ontological paradox'. Cf. [1] (p. 149).

[20] See, for instance, [1,3,14]. There is an ongoing debate about whether aesthetic value comes from nothing with regards to an artwork caught in a (time travel) loop: see [32–34]. CEN loops are not limited to time travel scenarios, see for instance [35] (p. 58).

[21] A similar line of thinking occurs in [36] (p. 301), with Harry's confidence in casting the Patronus: Harry finds the confidence to cast the difficult spell because he'd 'already done it', but thanks to time travel, the casting he (his earlier self) remembered and the casting he (his later self) performed were one and the same.

[22] Alasdair Richmond raises the possibility that a Kantian noumenal self seems just as capable of 'willing' (atemporally) a closed causal chain as a linear one, which lends some credence to the idea that causal loops are no harder to explain sub specie aeternitatis than linear causal chains [38] (p. 102).

[23] More recently, Meyer [2] (p. 260) considered the explicability of causal loops in light of two interpretations of the principle of sufficient reason.

[24] As is the case in [31]. Divine foreknowledge may be a counterexample to this, if one thinks God is omniscient.

[25] This is a common trope in fiction, with vague, misleading or deceptive prophecies. Examples can be found in *Macbeth* ("need fear none of woman borne"); *The Return of the King* (Eowyn and the Witch-king of Angmar); *Buffy the Vampire Slayer* ("Prophecy Girl"); *Mostly Harmless* (Arthur Dent's end); *Pirates of the Caribbean: On Stranger Tides* (Blackbeard's death).

[26] Indeed, if Rennick [25] is right about the impact of foreknowledge on intention formation, these cases would be even rarer. See also [24] (§§6.1–6.3) on why agents might try to 'bilk' (change) the past.

[27] There is an argument similar in spirit in [24] (see especially p. 377 re dates on objects).

[28] The first-person/third-person distinction does not map on to the grammar employed. For instance, Oedipus, upon learning of the prophecy, has first-person foreknowledge—although when referring to 'Oedipus' we speak of him in the third person.

[29] And it seems less uncanny that beliefs should arise from nowhere, if indeed that's what's happening in SFP cases.

References

1. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
2. Meyer, U. Explaining causal loops. *Analysis* **2012**, *72*, 259–264. [CrossRef]
3. Hanley, R. No End in Sight: Causal Loops in Philosophy, Physics and Fiction. *Synthese* **2004**, *141*, 123–152. [CrossRef]
4. French, P.A.; Brown, C. Time Travel. In *Puzzles, Paradoxes and Problems*; St Martin's Press: New York, NY, USA, 1987.
5. Wasserman, R. *Paradoxes of Time Travel*; OUP: Oxford, UK, 2017.
6. Mellor, D.H. *Real Time*; Cambridge University Press: Cambridge, UK, 1981.
7. Mellor, D.H. *Real Time II*; Routledge: London, UK; New York, NY, USA, 1998.
8. Mellor, D.H. *The Facts of Causation*; Routledge: London, UK; New York, NY, USA, 1995.
9. Berkovitz, J. On Chance in Causal Loops. *Mind* **2001**, *110*, 1–23. [CrossRef]
10. Bourne, C. *A Future for Presentism*; OUP: Oxford, UK, 2006.
11. Dowe, P. Causal Loops and the Independence of Causal Facts. *Philos. Sci.* **2001**, *68*, S89–S97. [CrossRef]
12. Riggs, P.J. A Critique of Mellor's Argument against 'Backwards' Causation. *BJPS* **1991**, *42*, 75–86. [CrossRef]
13. Ismael, J. Closed Causal Loops and the Bilking Argument. *Synthese* **2003**, *136*, 305–320. [CrossRef]
14. Macbeath, M. Who Was Dr. Who's Father? *Synthese* **1982**, *51*, 397–430. [CrossRef]

15. Skzrypek, J. Causal Time Loops and the Immaculate Conception. *J. Anal. Theol.* **2020**, *8*, 321–343.
16. Hunt, D.P. Providence, Foreknowledge, and Explanatory Loops: A Reply to Robinson. *Relig. Stud.* **2004**, *40*, 485–491. [CrossRef]
17. Robinson, M.D. Divine providence, simple foreknowledge, and the 'Metaphysical Principle'. *Relig. Stud.* **2004**, *40*, 471–483. [CrossRef]
18. Zimmerman, D. *The Providential Usefulness of 'Simple Foreknowledge'*; Rutgers University: New Brunswick, NJ, USA, 2010; Available online: http://fas-philosophy.rutgers.edu/zimmerman/Providence.Simple.Forek.6.pdf (accessed on 8 September 2021).
19. Sophocles. King Oedipus. In *The Theban Plays*; Watling, E.F., Ed.; Penguin Classics: London, UK, 1947; pp. 25–70.
20. Rowling, J.K. *Harry Potter and the Order of the Phoenix*; Bloomsbury: London, UK, 2003.
21. Wachowski, L.; Wachowski, A. The Matrix, Script. 1996. Available online: http://www.imsdb.com/scripts/Matrix,-The.html (accessed on 8 September 2021).
22. Ehasz, A.; O'Bryan, J. The Fortuneteller. In *Avatar: The Last Airbender*; Nickelodeon: New York, NY, USA, 2005.
23. Nix, G. *Lirael: Daughter of the Clayr*; HarperCollins: Sydney, Australia, 2001.
24. Smith, N.J.J. Bananas enough for time travel? *BJPS* **1997**, *48*, 363–389. [CrossRef]
25. Rennick, S. Things mere mortals *can* do, but philosophers can't. *Analysis* **2015**, *75*, 22–26. [CrossRef]
26. Goldman, A. Actions, Predictions and Books of Life. *Am. Philos. Q.* **1968**, *5*, 135–151.
27. Craig, W.L. Divine Foreknowledge and Newcomb's Paradox. *Philosophia* **1987**, *17*, 331–350. [CrossRef]
28. Effingham, N. *Time Travel: Probability and Impossibility*; OUP: Oxford, UK, 2020.
29. Dennett, D. True Believers: The Intentional Strategy and why it works. In *Mind Design II: Philosophy, Psychology, Artificial Intelligence*; Haugeland, J., Ed.; MIT: Cambridge, MA, USA, 1997.
30. Pratchett, T. *The Colour of Magic*; Transworld: London, UK, 1983.
31. Heinlein, R.A. —All You Zombies—. *The Magazine of Fantasy and Science Fiction*, March 1959.
32. McCall, S. An Insoluble Problem. *Analysis* **2010**, *70*, 647–648. [CrossRef]
33. Caddick Bourne, E.; Bourne, C. The Art of Time Travel: A Bigger Picture. *Manuscrito* **2017**, *40*, 281–287. [CrossRef]
34. McAllister, J.W. Does Artistic Value Pose a Special Problem for Time Travel Theories? *Br. J. Aesthet.* **2020**, *60*, 61–69. [CrossRef]
35. Gott, J.R., III; Li, L.-X. Can the universe create itself? *Phys. Rev.* **1998**, 58. [CrossRef]
36. Rowling, J.K. *Harry Potter and the Prisoner of Azkaban*; Bloomsbury: London, UK, 1999.
37. Levin, M.R. Swords' Points. *Analysis* **1980**, *40*, 69–70. [CrossRef]
38. Richmond, A. On behalf of spore gods. *Analysis* **2017**, *77*, 98–104. [CrossRef]
39. Aibel, J.; Berger, G. Kung Fu Panda, Script. Available online: http://www.imsdb.com/scripts/Kung-Fu-Panda.html (accessed on 14 September 2015).

philosophies

MDPI

Article

Exterminous Hypertime

Nikk Effingham

Department of Philosophy, University of Birmingham, Birmingham B15 2TT, UK; N.Effingham@bham.ac.uk

Abstract: This paper investigates 'exterminous hypertime', a model of time travel in which time travellers can change the past in virtue of there being two dimensions of time. This paper has three parts. Part one discusses the laws which might govern the connection between different 'hypertimes', showing that there are no problems with overdetermination. Part two examines a set of laws that mean changes to history take a period of hypertime to propagate through to the present. Those laws are of interest because: (i) at such worlds, a particular problem for non-Ludovician time travel ('the multiple time travellers' problem) is avoided; and (ii) they allow us to make sense of certain fictional narratives. Part three discusses how to understand expectations and rational decision making in a world with two dimensions of time. I end with an appendix discussing how the different theories in the metaphysics of time (e.g., tensed/tenseless theories and presentism/eternalism/growing block theory) marry up with exterminous hypertime.

Keywords: time travel; exterminous hypertime; hypertime; non-Ludovician; growing block theory; eternalism; presentism; overdetermination; truth in fiction

Citation: Effingham, N. Exterminous Hypertime. *Philosophies* **2021**, *6*, 85. https://doi.org/10.3390/philosophies6040085

Academic Editor: Alasdair Richmond

Received: 29 July 2021
Accepted: 22 September 2021
Published: 13 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Imagine I time travel from 2021 to 1930 and assassinate Hitler. For this to be possible, a 'non-Ludovician' model of time travel must be true. One such model is 'exterminous hypertime' whereby the changes are possible because time has two dimensions. This paper examines that model, expanding on previous papers of mine on the topic [1] (pp. 73–90) [2]. (Note that this 'hypertemporal theory' is not the same as that discussed by Goddu [3–5], van Inwagen [6], and Wasserman [7] (pp. 90–94); for a discussion of the differences, see [1] (pp. 76–90) and note 4).

After an initial exposition of the theory (Section 1), this paper proceeds in three parts. Part one considers the causal links between hypertimes, what I call 'hypercausation' (Section 2). As Section 3 explains, contrary to the likes of Smith [8] (and my earlier self [1] (pp. 83–84)), exterminous hypertime has no problem concerning overdetermination.

This paper does *not* argue that our world is *actually* one of exterminous hypertime—after all, for all I know time travel is not even physically possible[1]. Rather, this paper is an exploration only of what is metaphysically possible.

You might wonder what the point of such an exploration is. In Section 4, I examine different possible laws which might govern exterminous hypertemporal worlds. To show why those laws are of interest, Section 5 discusses two applications. Firstly, time travellers in exterminous hypertemporal worlds avoid an obstacle to time travel that threatens certain alternative non-Ludovician theories. Secondly, it allows us to make sense of certain fictional stories that are prima facie metaphysically impossible. That completes the second part of this paper.

In the third and final part, which follows from the discussion about how to understand what agents in fictional stories say and do, I discuss how to understand expectation and rational decision making in a world with two dimensions of time (Section 6). I argue that agents are rational in worrying about the hyperfuture and rational to accordingly change how they act in light of what might hyperlater happen.

This paper ends with an appendix on how exterminous hypertime connects with various issues in the metaphysics of time, namely tensed vs. tenseless time, the ontology of time, and the open future.

1. The Basic Theory

1.1. Two Dimensional Time

Exterminous hypertemporal worlds have multiple dimensions of time. In this paper, I assume such worlds have just the two (elsewhere, though, I have discussed versions of the theory with three or more [2]).

To better understand exterminous hypertime's multiple dimensions of time, compare them to the multiple dimensions of space. Space has three dimensions. Pictorial representations of space correspondingly have three axes, the x, y, and z axes. We can similarly represent two temporal dimensions with two axes, one being the regular temporal axis and the other being the hypertemporal axis. The x, y, and z axes of a spatial diagram are such that every point along the x axis exists at every point along the y and z axes (and every y point exists at every point along the x and z axes, and similarly for every z point). The same applies to exterminous hypertime. At every different hypertime, the entire complement of times exist. Use t_1, t_2 ... to name different instants of regular time (where t_n names an arbitrarily selected instant from n AD). Use T_1, T_2 ... to name different instants of hypertime. At T_1, every time exists, i.e., t_1, t_2... all exist at T_1. At T_2, every time also exists, i.e., t_1, t_2 ... all exist at T_2. And so on, for every hypertime.

Use the following notation to pick out different points of time-hypertime: 't_n-T_m' picks out time t_n at hypertime T_m, e.g., t_{1930}-T_1 picks out an instant from 1930 at hypertime T_1 whilst t_{2021}-T_{14} picks out an instant from 2021 at hypertime T_{14}.

1.2. Time Travel

I have previously argued [1] (p. 77) [2] that this two-dimensional picture of time can be used to understand time travel, as have others [7,9] (pp. 98–99) (see also [10]). This subsection summarises how this works.

Imagine I use a time machine to travel from 2021 to 1930 and successfully kill Hitler, preventing World War II. Exterminous hypertime can make sense of this scenario if we assume:

PROGRESSION: For any hypertime T_j, any regular time t_n, and any time traveller, x: if, at T_j, x time travels from t_n to $t_{m<n}$ then x travels to $t_{m<n}$-T_{j+1}[2,3].

Given PROGRESSION, time travellers move progressively forwards in hypertime. They can travel back in regular time, but never back in hypertime. When I kill Hitler, I start at (e.g.,) hypertime T_1. At T_1, in 2021, I activate my time machine and travel back to 1930. Given PROGRESSION, I leave T_1 and travel to hypertime T_2, arriving in the past in 1930. That is: I leave t_{2021}-T_1 and travel to t_{1930}-T_2. At T_1, history is such that Hitler survived 1930 and World War II started in 1939, i.e., Hitler is alive at t_{1930}-T_1 and war breaks out at t_{1939}-T_1. At T_2, history is different because I have assassinated Hitler. Having killed Hitler in 1930, he is dead at t_{1930}-T_2. This further prevents the outbreak of war, ensuring t_{1939}-T_2 is peaceful and harmonious. See Figure 1.

This theory is not without its problems. Questions include: What is hypertime exactly? Even though space can have multiple dimensions, is it not mind-boggling to think time can likewise have multiple dimensions? [7] (p. 99) [11] (p. 9) [12] (p. 209). Are the people at other hypertimes just *duplicates* of people at hyperearlier hypertimes—that is, do I not just end up killing a duplicate of Hitler when I kill the 'Hitler looking person' at t_{1930}-T_2? [13]. Is it accurate to call my making T_2 different from T_1 a case of 'changing the past'? [14,15] (p. 152) [16] (pp. 365–366) (see also [5]).

These are all good questions. However, I have dealt with them elsewhere [1] (pp. 79–90) and they are not the subject of this paper. Instead, this paper focuses on questions concerning the causal connections between the hypertimes.

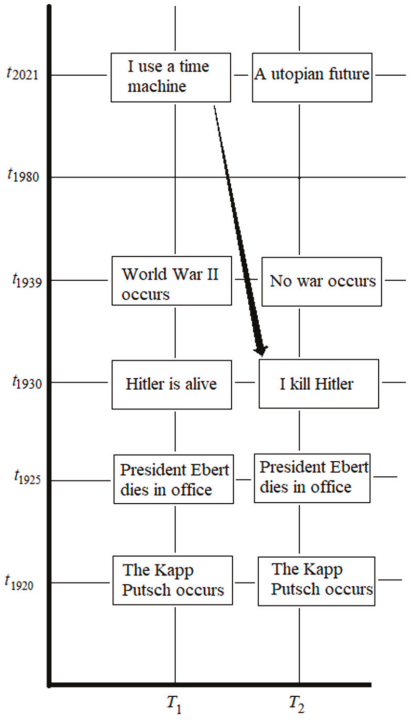


Figure 1. Exterminous hypertime and time travel.

2. Hypercausation

2.1. Stability

Not every world with multiple temporal dimensions is a world of exterminous hypertime. Both philosophers ([17,18] (p. 278) [19–22] (pp. 30–34) [23] (pp. 46–49) [24–26] (pp. 20–27) [27,28]) and physicists [29–35] have discussed worlds with multiple temporal dimensions, but such worlds are not anything like those discussed in this paper. More things must therefore be true to characterise worlds of exterminous hypertime. PROGRESSION is one such example. Another example would be:

> STABILITY: For all n and m, and for some j and k: t_n-T_m is qualitatively identical to t_n-T_{m-1} except when a time traveller from $t_{j\geq n}$-T_{m-1} has time travelled to $t_{k\leq n}$-T_m.

STABILITY ensures that the past does not change except when time travel occurs—left unmolested by time travellers, the past is always the same as it was at the immediate hyperearlier hyperinstant. For instance, at a world where no time travel takes place, then STABILITY entails that t_{1930} is the same at every hypertime, as is t_{2021}, and, indeed, all times (see Figure 2); whereas, in a world where time travel takes place, time-hypertimes may differ, but only if they are later in time than the arrival of some time traveller from a hyperearlier hypertime. Look back at Figure 1: every instant prior to t_{1930} is the same at T_1 and T_2, and it is only after my arrival at t_{1930}-T_2 that the time-hypertimes differ because I have killed Hitler.

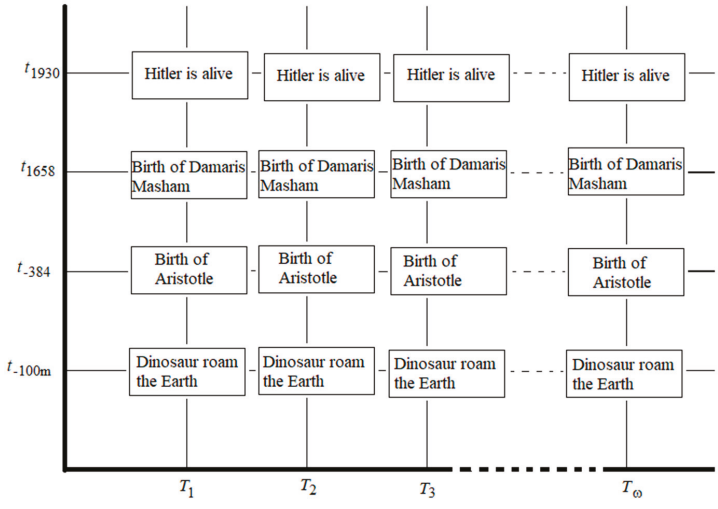


Figure 2. An exterminous hypertemporal world with no time travel.

STABILITY is important. In a one-dimensional temporal world, later times are suitably similar to earlier times. It is only via causal activity that the later times differ from the earlier times. STABILITY is one way of ensuring that a similar claim is true of hypertime. Given STABILITY, hyperlater hypertimes differ from hyperearlier hypertimes only because of the causal activity of time travellers.

Indeed, without STABILITY, it would be hard to recognise moving to hyperfuture hypertimes as being a form of time travel. Imagine STABILITY was false. Given it is false, historical events can play out differently at the different hypertimes, independently of how the world hyperearlier was. Imagine that: at T_1, history as we know it plays out; at T_2, the Big Bang is instead a Small Splutter and for all eternity the universe is an empty void; at T_3, the Big Bang occurs but history plays out differently and, by 2021, squid-headed aliens rule a universe composed of planets and galaxies foreign to those found at T_1. At T_1, in 2021 I build a time machine. I gather together my friends and, intending to amaze them, I travel back in time twenty seconds. Rather than appearing in the past surrounded by my companions watching me step into my time machine, I appear at T_2 in the empty void of space. Confused, I then decide to head back to 1930 to kill Hitler. I appear at T_3, surprised to find myself at the mercy of cephalopodic monsters who live in a galaxy that did not even exist when I was younger. Fantastic things are happening when I activate my machine, but travelling back in time does not seem to be one of them. Twenty seconds ago, the past was full of material objects, so how can I have travelled back twenty seconds if the place I arrive at it is devoid of all matter? Ninety one years ago, the Weimar Republic existed, so how can I have arrived in 1930 if the Weimar Republic is instead nowhere to be found? In an unstable world, movement between hypertimes does not seem to give rise to what we would normally think of as 'time travel' [36] (pp. 379–380). Given STABILITY, this problem is avoided. Hence, we should endorse STABILITY.

2.2. Explaining Stability

Whilst STABILITY might be inexplicably true, it is less strange to think that certain laws of nature will explain it. The rest of this paper assumes that STABILITY is true because certain hyperearlier time-hypertimes stand in a causal relation to other hyperlater time-hypertimes. Dub that causal connection 'hypercausation'. (Note that hypercausation is not a 'metaphysically different' type of causation; the introduction of the term is solely for the purpose of exposition.)

Given such hypercausal connections, we can explain STABILITY. When I travel back to t_{1930}-T_2, that time-hypertime is such that Hitler is alive (and the Weimar Republic is running Germany, etc.) because the immediately hyperprior time-hypertime (i.e., t_{1930}-T_1) hypercausally influences t_{1930}-T_2 in a way that makes it almost entirely qualitatively identical to t_{1930}-T_1, differing only with regard to my arrival in a time machine.

3. Overdetermination

Some people have worried that there being hypercausal connections between time-hypertimes gives rise to a problem of systematic overdetermination. Section 3.1 explains the worry. Section 3.2 argues that, even if there were systematic overdetermination, it is not worrying. Section 3.3 argues that, in any case, we need not believe that there is such systematic overdetermination in the first place.

3.1. The Threat of Overdetermination

An event may be a 'partial' cause of its effect or a 'whole' cause. For example, if three people lift a table then each person's lifting is a partial cause of the table being lifted. It is only their collective action that wholly causes the lifting.

Given that causation is transitive, effects can have multiple whole causes. If e_1 wholly causes e_2 which in turn wholly causes e_3, then e_1 and e_2 are both whole causes of e_3. Such cases are routine and occur all of the time. But some cases of effects having multiple whole causes are weirder. In the routine cases (e.g., of e_1 and e_2 causing e_3) we can trace a single chain of causes threading all three events. In the weird cases, an effect has multiple whole causes from *separate* causal chains. For instance, imagine two assassins independently, yet simultaneously, kill a target, e.g., an assassin's bullet strikes the victim in the heart at exactly the same moment that the second assassin's poison causes a terminal encephalopathic event. This is weird in a way that the chain of e_1, e_2, and e_3 is not weird. Call the weird cases 'overdetermination cases'.

Whilst weird, overdetermination cases are possible. Yet, we tend to think they do not *regularly* take place, i.e., we think *systematic* overdetermination is not taking place. (As a forewarning, Section 3.2 questions exactly this assumption). If STABILITY is explained hypercausally, then there is at least a prima facie reason to think systematic overdetermination will be a problem—a worry previously noted by both Smith [8] (pp. 691–692) and myself [1] (pp. 83–84).

To see why we might suspect that there is systematic overdetermination, imagine a world at which there is no time travel and consider the 1804 liberation of Haiti. At t_{1804}-T_1, the self-liberated slaves have finally overthrown French rule. Assuming there is no hypertime prior to T_1, t_{1804}-T_1 is the way it is solely because, at earlier (regular!) times at T_1, the slaves did things to cause the liberation at t_{1804}-T_1. Since t_{1804}-T_2 is qualitatively identical to t_{1804}-T_1 (since all earlier times at T_2 are qualitatively identical to the earlier times from T_1), the liberation of Haiti at t_{1804}-T_2 is likewise wholly caused by the actions of slaves from earlier times at T_2. However, given the hypercausal explanation of STABILITY, the liberation of Haiti at t_{1804}-T_1 *also* wholly causes the liberation of Haiti at t_{1804}-T_2. We have a case of overdetermination! And what goes for the liberation of Haiti goes for every event at every hypertime other than T_1. So, at such a world, there is systematic overdetermination. Given the assumption that there is no systematic overdetermination, exterminous hypertime would have a problem[4].

3.2. There Is No Problem of Overdetermination

However, this is much too quick. Firstly, as already noted, overdetermination cases are possible, even if they are unlikely. An anonymous referee (for my monograph [1], not this paper) correctly pointed out that these worries about systematic overdetermination only show that we probably do not *actually* live in an exterminous hypertemporal world. Since this paper is interested only in what is possible—and readily admits that we do not actually

live in an exterminous hypertemporal world—then worries about overdetermination are by-the-by.

Secondly, if the problem of systematic overdetermination is that it is unlikely, then it is not a problem in any case[5]. When we say systematic overdetermination is 'unlikely', the type of likelihood in question is 'objective chance'. Objective chance is intimately tied up with the laws of nature. However, at a world of exterminous hypertime, the systematic overdetermination is *mandated* by the laws of nature (see Section 4 for examples of such mandating laws). Since it is mandated, the overdetermination is therefore *not* unlikely. Indeed, since the laws mandate it, it always has a chance of occurring equal to 1—the overdetermination is *as likely as it could possibly be*. The liberation of Haiti at t_{1804}-T_2 may be wholly caused by two distinct events, but it is no coincidence! Thus, even if there is systematic overdetermination, it is not of the problematic kind[6].

One objection might be that the *laws themselves* are unlikely, i.e., it is unlikely to find oneself in a world with laws mandating systematic overdetermination. However, whilst debating the objective chance of an event given certain laws is straightforward, discussing the likelihood of the laws themselves is not. We are immediately dragged into the quagmire of issues involved in the fine-tuning debate (see, e.g., [37,38] (pp. 147–154) and [39]). Since this objection quickly becomes stuck on such issues, I set it aside.

3.3. Exterminous Hypertemporal Worlds Need Not Be Worlds of Systematic Overdetermination

Section 3.2 argued that if an exterminous hypertemporal world contained systematic overdetermination then we need not worry. This section argues that we do not even need to go that far, because we need not think that exterminous hypertemporal worlds feature systematic overdetermination in the first place.

Introduce another chance function alongside the objective chance function. Call it the 'counterhypercausal chance function'—or '$\text{chance}_{\text{chc}}$', for short. The $\text{chance}_{\text{chc}}$ of an event occurring (at some time t) is the objective chance it would have of occurring (at t) if only nothing hypercausally interfered with it. As an example, consider the following case:

> *Case one: Moscovium decay with no time travel*. At t_{2021}-T_1 a moscovium atom has 0.5 chance of decaying a few millseconds later. And, indeed, it decays. Further, no time travel takes place at this world. At t_{2021}-T_2, consider the same atom. Since no time travel has occurred, it has a $\text{chance}_{\text{chc}}$ of decaying equal to 0.5. However, given Stability it must decay, so at t_{2021}-T_2 it has a chance of decaying equal to 1[7].

This in place, turn back to how a world of exterminous hypertime can avoid systematic overdetermination. The key is to assume that a time-hypertime hypercausally influences another time-hypertime *only* regarding probabilistic events, i.e., events not determined to happen. In such cases, the hypercausal interaction either ensures that the event happens (if it occurred at the hyperprevious time-hypertime) or ensures that the event does not happen (if it did not occur).

Return to the Haiti example. Imagine that the laws of nature are such that the actions of the self-liberated slaves at t_{1803}-T_2 determine most of what goes on at t_{1804}-T_2, i.e., most events at t_{1804}-T_2 are non-probabilistic. At t_{1803}-T_2, the remaining events at t_{1804}-T_2 have a $\text{chance}_{\text{chc}}$ of occurring between 0 and 1. It is those remaining events (and only those events) that are influenced by hypercausation. Thus, the earlier events of t_{1803}-T_2 are only a partial, not whole, cause of the liberation of Haiti. Similarly, the events of t_{1804}-T_1 are only a partial, not whole, cause of the liberation at t_{1804}-T_2. It is only the *conjunction* of these things that is the *whole* cause. Therefore, there is no overdetermination (and so no systematic overdetermination) because overdetermination involves distinct *whole* causes bringing about some effect, not distinct *partial* causes bringing about some effect. Problem solved.

4. Hypercausal Laws

This section examines the different hypercausal laws that would allow for a world of exterminous hypertime. Sections 2 and 3 have argued that all such laws must entail STABILITY; however, there are other dimensions along which they may vary.

4.1. Local vs. Global

The first possible hypercausal law for consideration is:

GLOBAL EFFECT: For any event, e, and all j, k, m and n, if e occurred at t_m-T_{n-1} then whether or not it occurs at hyperlater hypertimes depends upon whether a time traveller has (or has not) time travelled from $t_{j\geq m}$-T_{n-1} to $t_{k\leq m}$-T_n:

> If a time traveller has *not* time travelled from $t_{j\geq m}$-T_{n-1} to $t_{k\leq m}$-T_n then e occurs iff it occurred at t_m-T_{n-1} (except where t_m-T_{n-1} does not exist—i.e., where T_n is the first hypertime—in which case e has a chance of occurring equal to its $\text{chance}_{\text{chc}}$ of occurring).
>
> If a time traveller *has* time travelled from $t_{j\geq m}$-T_{n-1} to $t_{k\leq m}$-T_n then e may or may not have occurred. Its occurrence will be probabilistically governed; the chance (at t_l-T_n, for all $l < m$) of e occurring is equal to its $\text{chance}_{\text{chc}}$ of occurring[8].

Given GLOBAL EFFECT, time-hypertimes are qualitatively identical to their immediately hyperprevious-yet-simultaneous time-hypertimes unless someone has time travelled to that time or earlier. And if a time traveller has travelled to that time or earlier, then all hypercausal ties are severed—from that point forwards, 'all bets are off' when it comes to how the future can play out, with no hypercausation affecting matters.

To see the upshot of GLOBAL EFFECT, consider four example cases. Start with Case One from above. In Case One, the moscovium atom decayed just after t_{2021}-T_1. Since there is no time travel in Case One, GLOBAL EFFECT entails that it also decays just after t_{2021}-T_2 (Which is just as it should be, because this is just what STABILITY requires).

Next, consider:

> *Case Two: Moscovium decay with time travel and interaction*. As per Case One, but a time traveller from t_{3000}-T_1 has time travelled to T_2 to an instant a few hours earlier than t_{2021}. There, the time traveller accelerates the moscovium atom to close to the speed of light, lowering its $\text{chance}_{\text{chc}}$ of decaying from 0.5 to 0.1.

Given GLOBAL EFFECT, because the time traveller has arrived at a time earlier than t_{2021}-T_2, the chance (at t_{2021}-T_2) of the atom decaying a few milliseconds after t_{2021}-T_2 ends up being independent of the hyperprior hypertime, i.e., the chance is 0.1 (rather than, as in Case One, a chance of 1). So, it may very well *not* decay, quite unlike with Case One.

> *Case Three: Moscovium decay with time travel but no interaction*. As per Case Two, but the time traveller does not interact with the moscovium atom (for instance, because they arrive on the other side of the world).

Even though the time traveller does not interact with the moscovium atom, given GLOBAL EFFECT, the hyperearlier time-hypertime nevertheless no longer hypercausally affects t_{2021}-T_2. That means that (at t_{2021}-T_2) the chance of the atom decaying is 0.5; at t_{2021}-T_2 the moscovium is as liable to decay as not.

Finally, consider:

> *Case Four: The Andromeda case*. At t_{2021}-T_1 there exists a time portal. Stepping through it, you travel to the Andromeda galaxy in 50,000 BC. Staying for a few minutes, you then use the portal to return back to the future on Earth. Obviously, what you do in Andromeda's Palaeolithic past does not affect events on the planet Earth in its history leading up to the present day.

Given GLOBAL EFFECT, you will almost certainly find history has played out differently. Imagine that an ancestor of Aristotle from 40,999 BC did not die of cancer before

reaching child-bearing age at T_1 because of an unlikely stochastic event. At T_2, your arrival in Andromeda means that the ancestor's survival is again put in doubt. Imagine that what is most likely happens and the ancestor *does* die. Resultantly, Alexander the Great fails to learn the correct lessons from Aristotle, never defeats the Acaehmenid Empire, and history is radically changed. Even though you spent only a minute or two in the Andromeda galaxy, when you return through the time portal to 2021, you find that the world is now under the control of the NeoPersian Hegemony. Everyone you know has gone; your ancestors of the previous thousands of years never existed; Earth's history is nothing like that which you remember.

So, as Case Three and Case Four make clear, GLOBAL EFFECT means that *any* time travel interferes with history on a universal scale. Even things far removed from a time traveller can be affected by their act of time travel.

GLOBAL EFFECT is not the only possible law that there might be at a world of exterminous hypertime. Consider:

> LOCAL EFFECT: For any event e, and all m, n and k, whether an event e occurs at t_m-T_n depends upon whether someone has or has not time travelled from $t_{j \geq m}$-T_{n-1} to $t_{k \leq m}$-T_n:
>
> > If a time traveller has *not* time travelled from $t_{j \geq m}$-T_{n-1} to $t_{k \leq m}$-T_n then e occurs iff it occurred at t_m-T_{n-1} (except where t_m-T_{n-1} does not exist—i.e., where T_n is the first hypertime—in which case e has a chance of occurring equal to its $\text{chance}_{\text{chc}}$ of occurring).
> >
> > If some time traveller *has* time travelled from $t_{j \geq m}$-T_{n-1} to $t_{k \leq m}$-T_n then e's occurring or not depends upon whether some time traveller has causally interacted/influenced any factors governing e's occurring or not. If some such factor has been influenced by a time traveller, then e's occurring is now independent of what went on at the hyperearlier time-hypertime; e's chance of occurring is equal to its $\text{chance}_{\text{chc}}$ of occurring. If some such factor has not been influenced by a time traveller (e.g., the time traveller arrived too spatiotemporally distant to influence e's coming about or not) then e occurs iff it occurred at t_m-T_{n-1}.

To see the difference between LOCAL EFFECT and GLOBAL EFFECT, reconsider the cases. Cases One and Two play out exactly as they do given GLOBAL EFFECT. However, in Cases Three and Four, LOCAL EFFECT and GLOBAL EFFECT diverge. Consider Case Three. Arriving in the past, the time traveller does not causally interact with the moscovium atom. Given the second clause of LOCAL EFFECT, e must therefore occur, contrary to what was true given GLOBAL EFFECT. Similar thoughts apply to Case Four. Arriving in Andromeda, the time travellers cannot causally influence events in the Milky Way. Hence, events in Earth's past are still held tight in the grip of hypercausation, playing out exactly as they did at T_1. Given LOCAL EFFECT, when the time travellers return to the future, they find everything exactly as it was before.

So exterminous hypertime might allow for time travel having a local or global influence. If it is global, then as soon as a time traveller arrives in the past, the entire universe's future is unshackled from slavishly following the course of the previous hypertime. If it is local, only those events that the time traveller interacts with can turn out differently from the previous hypertime.

4.2. Propagative Laws

The local/global dimension is not the only dimension which the laws of nature can vary over. Both LOCAL EFFECT and GLOBAL EFFECT have it that when the future is changed, it is changed 'hyperinstantaneously', i.e., when a time traveller intercedes in history, the very next hyperinstant is such that the entire future reflects the changes that have been made. However, we might instead believe that the changes to history 'propagate', taking a period of hypertime to change the future. That is: A time traveller changes the past, but at

the next hyperinstant those changes are only reflected at a tiny bit of history; as hypertime moves on, more and more of history changes to reflect the time traveller's actions.

As an initial example, imagine that I go back and kill Hitler. If changes propagate, then the next hyperinstant of hypertime will not reflect that change. Even though, at the next hyperinstant, I killed Hitler in 1930, he is still alive in 1939 and starting World War II. We have to wait until a hyperlater hypertime for history to have changed such that Hitler is now dead in 1939 and World War II does not occur. That is: The changes caused by time travellers propagate over a *hypertemporal* period.

This subsection details the laws required to make exterminous hypertime 'propagative'. (I will assume throughout that time travel has a global influence, similar to GLOBAL EFFECT, though I see no reason to think that a local version, riffing off of LOCAL EFFECT, could not instead be constructed).

First, we must say that there is a property of 'hyper-resilience'. It is had (or not) by time-hypertimes. Three hypercausal laws govern hyper-resilience:

1. For all j and k, t_j-T_k is *not* hyper-resilient iff something tried to causally influence t_j-T_{k-1} that did not try to causally influence t_j-T_{k-2}; otherwise, t_j-T_k is hyper-resilient.
2. For all j and k, if t_j-T_k is hyper-resilient then events occur at it iff they occurred at t_j-T_{k-1}. This is the case even if something (e.g., a time traveller coming from the previous hypertime) tries to causally influence them to be otherwise, i.e., hyper-resilience 'trumps' all other forms of causal force and activity.
3. If a time-hypertime is *not* hyper-resilient then there is no hypercausal influence between its hyperprior time-hypertime and it. What goes on at such time-hypertimes is governed only by what went on at hypersimultaneous earlier times in combination with the causal influence of any time traveller who has arrived from the hyperprevious hypertime.

Call the conjunction of these claims GLOBAL PROPAGATION. If GLOBAL PROPAGATION is a law of nature, then changes to history take a period of hypertime to issue forth through to the present. Consider how killing Hitler would pan out given GLOBAL PROPAGATION. Figure 3 depicts what would happen: time-hypertimes are depicted as the larger boxes; whether or not they are hyper-resilient is represented by a tic/cross in the smaller box, where a tic represents that the time-hypertime is hyper-resilient and a cross indicates that it is not. In accord with law (1), every time-hypertime at T_1 is hyper-resilient. At T_1 I attempt to travel from T_1 to 1930 to kill Hitler. Given law (1), every time-hypertime at T_2 is also hyper-resilient. So, given (2), every time-hypertime at T_2 is the same as it was at T_1; since I was not in 1930 at T_1 then, even though I *try* to travel to t_{1930}-T_2, the hyper-resilience of t_{1930}-T_2 trumps my attempt; thus, I do not appear at t_{1930}-T_2. However, no time traveller tried to causally interact with t_{1930}-T_1 by trying to arrive there, whilst a time traveller *does* try and causally interact with t_{1930}-T_2 by trying to arrive there. So, given (1), t_{1930}-T_3 is *not* hyper-resilient. Since it is not hyper-resilient then, given (3), events at t_{1930}-T_3 can play out differently than t_{1930}-T_2. So, whilst I do not arrive at t_{1930}-T_2 from t_{1930}-T_1, I *do* arrive at t_{1930}-T_3 (having arrived there from t_{2021}-T_2). However, I fail to persist any further at T_3. This is because the next time-hypertime in T_3, call it $t_{1930+\delta}$-T_3, *is* hyper-resilient; given (2), it must be qualitatively identical to $t_{1930+\delta}$-T_2; since I did not exist at $t_{1930+\delta}$-T_2, then I do not exist at $t_{1930+\delta}$-T_3 either. However, because something tries to causally affect $t_{1930+\delta}$-T_3 that did not try to causally affect $t_{1930+\delta}$-T_2, $t_{1930+\delta}$-T_4 is *not* hyper-resilient. So, I arrive at t_{1930}-T_4 and *can* persist to $t_{1930+\delta}$-T_4. I cannot, however, persist to the instant after that. At T_5, the same reasoning means that I can manage to persist to the instant after, but no further. At T_6, similar reasoning dictates that I manage to persist one instant more. And so on. Ultimately, we arrive at a hyperinstant, T_ω, where my interaction with the past has hyperfinally influenced the future, stopping World War II and bringing about a utopian future[9].

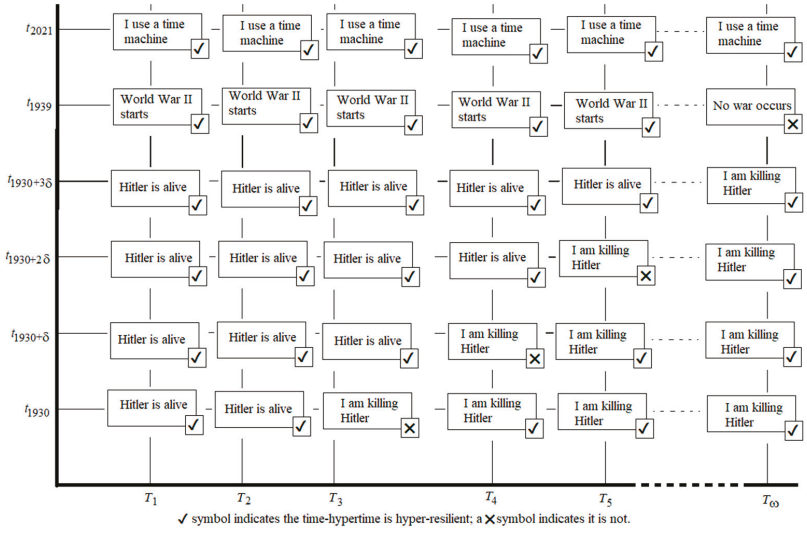


Figure 3. A globally propagative time travel scenario.

5. Applications of Global Propagation

You might wonder why I care about worlds at which GLOBAL PROPAGATION is true. After all, lots of weird things are metaphysically possible, so why be interested in *this* weird possibility? This section explains two interesting features of propagative worlds.

5.1. The Multiple Time Travellers Problem

Some theories that allow for the past to change face the 'Multiple Time Travellers Problem'. Versions of this problem have been discussed by Hudson and Wasserman [9] (p. 42) (see also [7] (p. 96 n. 73) as well as myself [36,40]. This section details that problem and then explains how GLOBAL PROPAGATION avoids it.

Start by imagining that changes to the past do *not* propagate through to the future and that GLOBAL EFFECT is true. At t_{2021}-T_1 I decide to travel to the past to kill Hitler. However, Malcolm—who is at t_{2022}-T_1—also uses a time machine. Malcolm travels to 50,000 BC (at T_2), kills Aristotle's ancestor, and brings about the NeoPersian Hegemony. Given PROGRESSION, we both arrive at T_2 (albeit at different times: me arriving in 1930 and Malcolm arriving in 50,000 BC). At t_{1930}-T_2, I arrive and expect to find Hitler and Germany but, because of Malcolm's interference, I instead find that Germany, Hitler, and all of Hitler's ancestors for a thousand generations never existed. As discussed in Section 2.1, if a putative time traveller arrives in the past and does not find such things, that is problematic because they do not appear to really have a time machine. So we have a problem. Call it the 'Multiple Time Travellers Problem': At a world where GLOBAL EFFECT is true, if multiple time travellers go back to the past, only those time travellers arriving at the earliest moment manage to succeed; everyone else arrives somewhere which is not properly seen as being 'their past' and so fails to travel in time.

GLOBAL PROPAGATION avoids this problem[10]. In a world where GLOBAL PROPAGATION is true, neither myself nor Malcolm arrive at T_2. However, we do arrive at T_3, if only for an instant. Malcolm's actions at $t_{-50,000}$-T_3 do not have any causal effect on events in the (regular) future, so when I arrive at t_{1930}-T_3, it is exactly as I expect it to be, i.e., with Germany, the Weimar Republic, and Hitler all in existence. So I *do* arrive somewhere properly called 'the past'. Thus, I manage to travel in time even though Malcolm *also* manages to travel in time.

As hypertime moves on, more and more of the past will be affected by our changes. For instance, there will be a hypertime, T_ω, at which ten years of time apiece have been 'affected'. At T_ω:

- 50,000 BC–49,990 BC include Malcolm's arrival in the past and the changes he has made;
- 49,990 BC–1930 AD are just as they are at T_1 (i.e., with no Malcolm at them, nor affected by anything Malcolm did at 49,990 BC or earlier);
- 1930 AD–1940 AD include my arrival in the past, assassinating Hitler, and preventing World War II;
- 1940 AD onwards is exactly as it originally was at T_1, i.e., Hitler is alive, World War II is taking place, etc.

If we hyperwait long enough, then we get to a hyperlater hypertime, $T_{\omega'}$, where Malcolm's interaction with the past finally catches up to 1930. At $T_{\omega'}$, I arrive in 1930 (having come from t_{2021}-$T_{\omega'-1}$) and am bewildered to find Germany has gone, the Weimar Republic does not exist, Hitler was never born, etc., and all such things have been replaced by the NeoPersian Hegemony. In that case I have *not* time travelled from $T_{\omega'-1}$ to $T_{\omega'}$. Nevertheless, GLOBAL PROPAGATION salvages the possibility of me managing to time travel back to 1930 for at least *some* period of hypertime (namely all of the hypertimes between T_1 and $T_{\omega'}$).

Thus, we have at least one reason to take note of worlds at which GLOBAL PROPAGATION is true: At such worlds, multiple time travellers can go back in time to different destinations.

5.2. The Metaphysical Possibility of Time Travel Fictions

Recognising the possibility of propagative worlds also helps with determining which time travel stories are/are not metaphysically possible. This was Lewis's task in his famous contribution to the philosophy of time travel [15]. He demonstrates that various time travel stories are possible (such as stories by Heinlein). Lewis was clear, though, that not any old time travel story was thereby possible and it remains interesting to ask of other stories whether they are possible or not. (Elsewhere [2], I have already discussed the possibility of other fictions, not covered by Lewis's theory, nevertheless being possible—this paper furthers that investigation).

GLOBAL PROPAGATION allows us to capture the metaphysical possibility of it 'taking time' for changes to the past to 'reach' the present. This trope appears in a variety of stories, such as Baxter's *Timelike Infinity* [41], Mark Millar's *Chrononauts* [42], Robert J. Sawyer's 'On the Surface' [43], *Star Trek: Enterprise*'s 'Carpenter Street' [44], and *Red Dwarf*'s 'Timeslides' [45], to name just a few. However, as this section explains, it turns out that only some such fictions turn out to be possible given GLOBAL PROPAGATION.

Consider a narrative that GLOBAL PROPAGATION does *not* bear out the possibility of: *Red Dwarf*'s 'Timeslides' [45]. Because of misadventure, the protagonist, Lister, is the last man alive in 3,000,000 AD. Finding a time machine, he returns to when he was a young adult (say, 2174 AD), changing history so he avoids his fate. He then returns to the future. At first, nothing changes. Asking why the past has not altered, Lister is informed that 'it'll take a few seconds for the timelines to sort themselves out'. Sure enough, a few seconds later the changes to history catch up to the present and Lister ceases to be; history is now different, with him having now lived a successful life in the past.

Whilst 'Timeslides' involves changes to the past taking a while to reach the present, its narrative is nevertheless *not* possible at a propagative world of exterminous hypertime. See Figure 4. At the earliest hypertime, T_1, the past is such that Lister lives an unsuccessful life and goes back in time to change this. At T_ω, the past up until 2175 AD has changed. At $T_{\omega'}$, still more of the past has altered; Lister is now rich and famous in 2180 AD, rather than poor and unknown. At $T_{\omega''}$, even more of the past has changed and history is different up until 3000 AD. And so on, until we get to a hypertime, $T_{\omega'''}$, at which all of history reflects the changes Lister made. However, consider what happens at each hypertime. At $T_{\omega'}$, Lister returns to the future and ... keeps on living. He does not vanish or fade away. Similarly for $T_{\omega''}$ and all other hypertimes hyperprior to $T_{\omega'''}$. Additionally, at $T_{\omega'''}$ there

is no Lister going back in the past in the first place because he lived a successful life and never got trapped in deep space. So there is *no* hypertime at which Lister stands around waiting to be washed away by the changes he made to the past, only to have those changes 'catch up with him'. 'Timeslides' is *not* the sort of narrative which GLOBAL PROPAGATION bears out[11].

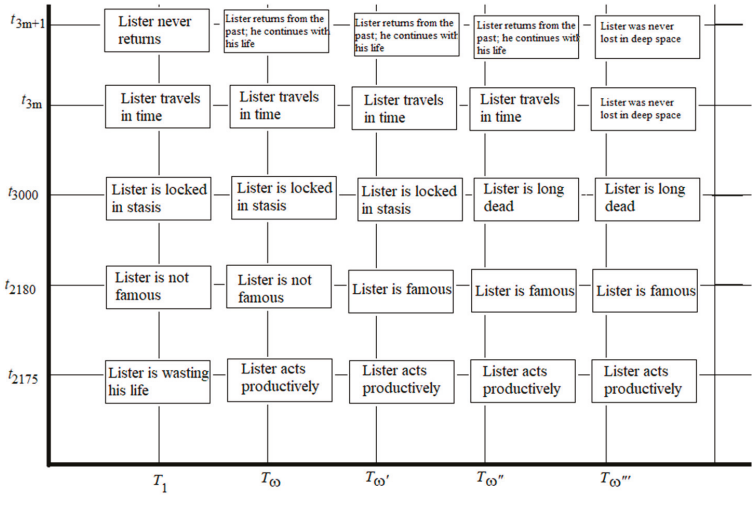


Figure 4. The narrative of *Red Dwarf*'s 'Timeslides'.

However, that does not mean that propagative worlds cannot capture *any* narrative that involves changes to history taking time to ripple forwards. Given a suitably liberal interpretation, some such fictions are possible given GLOBAL PROPAGATION. Consider two examples.

The first example is Baxter's *Timelike Infinity* [41]. Rebels have headed into the past to change history. One character, Jasoft Parz, wonders how their changes will affect him. He worries that there are no 'estimates of the rate—in subjective terms—at which the disruption was approaching, emerging from the past as if from the depths of some dismal sea' [41] (pp. 182–83). Whilst we might interpret Parz's worry as reflecting the idea that changes to history are propagating through regular time, there is a legitimate interpretation whereby his worry is well-founded because the disruption is approaching hypertemporally. At the current hypertime, Parz will be unaffected. But Parz knows—and is concerned about—what hyperwill happen to him, and whether the changes to the past will affect him at some hyperlater hypertime (Such an interpretation requires it to be rational for Parz to worry about what will happen to him at hyperlater hypertimes; I pick this issue up in Section 6).

The second example is Millar and Murphy's *Chrononauts* [42]. The characters worry that changes in the past are coming towards them, threatening to radically alter how things are. One character asks 'How long have we got?', asking how much longer they have, in the present, before the changes wipe them out.

It is natural to read this as a worry about changes to the past propagating through regular time. However, again, there is a legitimate interpretation whereby the changes are propagating through hypertime instead. Given that interpretation, we read the question 'How long have we got?' as concerning what period of *hypertime* it takes for the changes to reach the present. Given that interpretation, the comic still makes sense. And—assuming it is rational to worry about disasters in one's hyperfuture—there is little reason *not* to interpret the strip that way.

6. Hyperexpectations

Section 5.2 argued that certain fictions are possible given GLOBAL PROPAGATION and assumed that it is rational to worry about what happens to you in the hyperfuture. This section argues for that assumption by discussing connected issues concerning persistence, expectation, and rational decision making in a world with two temporal dimensions.

6.1. Personal Hyperfutures

Set aside the question of Jasoft Parz and the Chrononauts for the time being. Consider instead the following case:

> *Case Five. Survival in a propagative world*. *Global Propagation* is true. I am wondering whether to time travel from t_{2021}-T_1 to t_{1930}-T_2, i.e., I am wondering whether to time travel to 1930 AD. If I did, then the world would be as depicted in Figure 3. But—as Figure 3 makes clear—when I leave t_{2021}-T_1, I fail to appear at t_{1930}-T_2.

The worry is this: Since I fail to appear at t_{1930}-T_2, am I not effectively killing myself by travelling in time?

I believe this worry to be misplaced. To see why, consider a world with just a single dimension of time where I travel back in time from 2021 to the 1930s, intending to retire and never to return. We might worry that this effectively kills me, since—at 2021—I stop existing in the future. However, there is a well-rehearsed objection to this, relying on the distinction between me having an 'external future' and a 'personal future'—if you are unfamiliar with the distinction between external time and personal time, see [15] and [1] (pp. 42–45). In that I do not exist at 2022, 2023, 2024, etc., I do not have an external future. However, I still have a personal future lying in the 1930s. It is my having a personal future that is important to rational decision making, hence why my time travelling retirement plan is rational.

Whilst it is true that, in Case Five, when I activate my time machine at t_{2021}-T_1 I have neither an external future nor a personal future, this line of thinking nevertheless helps. Given exterminous hypertime introduces two dimensions of external time, we should do the same for personal time: I have two types of personal history, my regular personal history and my personal hyperhistory. Just as my future stages are those stages which I immanently causally influence, my hyperfuture stages are those that I immanently causally influence by how I am now and by what I hyperpresently do. So in Case Five, there is a stage of me in the hyperfuture (e.g., at t_{1930}-T_3, t_{1930}-T_3, t_{1930}-T_4 … t_{1930}-T_ω …) that I have immanently causally influenced, thus it is one of my personal hyperfuture stages[12]. So whilst it is true that I may have no more personal future, I nevertheless have a personal hyperfuture. And it is in that personal hyperfuture that I end up having a personal future where I activate the time machine and go on to survive in the 1930s.

An event lying in one's personal hyperfuture can be action guiding; my still having a post-time-machine-activation personal future at some hyperfuture time means that it is rational to travel back in time in Case Five. This is because it is rational to care about one's personal hyperfuture in the same way that one cares about one's personal future—your personal hyperfuture stages are stages of *you*, after all. Since they are stages you can still causally influence—in this case, hypercausally influence—it is natural to worry about what they hyperwill be like. Compare: In a world of one-dimensional time, it is rational for me to presently make sacrifices in order to benefit my future self. It is, for instance, rational to refrain from enjoying my delicious piña colada in order to flee a nearby erupting volcano, thus ensuring that my future self can drink many more cocktails later on. In a world of exterminous hypertime, the analogue applies and I should be motivated to make sacrifices to benefit my hyperfuture self, ensuring that my hyperfuture self has a personal history that lies in the (regular) past. Thus, if you want to travel back to the 1930s, it is rational to get into use the time machine in Case Five.

Indeed, we might be *more* motivated to worry about our hyperfutures than our regular futures. My biology is such that I am likely to live in the region of eighty or so years. However, given how hypercausation functions, I could potentially persist in the

hypertemporal direction for hypereternity. That said, time travellers might end up in a situation where they have to make hard choices. Imagine it is T_1 and a supervillain plans to go back in time and destroy all life on the planet from 50,000 BC onwards. If that happens at T_2, you will have existed only hypermomentarily. But imagine the supervillain makes you an offer: In return for not foiling his scheme, the villain promises you an elixir giving you an extended life span, with excellent health throughout, for four hundred years. What do you do? Which is better? A longer 'temporally regular' existence but at only one hypertime? Or a shorter 'regular' existence but at a potentially infinite number of hypertimes?

6.2. Hyperexpectations

Having bifurcated personal time, we should similarly bifurcate the notion of *expectation*. Consider the following case.

> *Case Six. The Retrorifle Case*. I live in a world where GLOBAL EFFECT is true. Rather than time travelling to 1930 in order to kill Hitler I instead fire a bullet from my 'retrorifle'. Such a gun fires a bullet back in time from t_{2021}-T_1 to t_{1930}-T_2, shooting Hitler dead. However, I myself do not travel in time; I remain at T_1.

In Case Six, what should I expect to happen when I fire the gun? At first glance, I should expect nothing to happen when I pull the trigger of the gun since my personal future lies wholly within T_1 and my retrorifle only causes history to be different at hyperlater hypertimes. So, we might think that I should not expect my firing the rifle to change the world around me.

There is something both substantially right about this, but also substantially wrong. Whilst my personal future will not reflect Hitler having being assassinated and World War II not happening, in my personal hyperfuture the world's history is such that Hitler is dead. Having bifurcated personal time, we should do the same for expectation: Whilst I should not *expect* anything to change when I pull the retrorifle's trigger, I should nevertheless *hyperexpect* things to change. My expectations track what will happen to me in my personal future; my hyperexpectations track what hyperwill happen to me in my personal hyperfuture. We can then reparse the claim that what happens to me in my hyperfuture is relevant to what decisions I make: What decisions I make should be based, not just on what I expect to happen given my actions, but also (at least partially) on what I hyperexpect to happen given my actions.

All that said, return to Jasoft Parz and the rationality of his worries. Given what I have said about expectations and hyperexpectations, Parz's worries are rational because he is correct to hyperexpect that he, and the world around him, will be different in the hyperfuture. Indeed, given the changes made to the past, he might not even exist at all once the changes hyperreach him—his chance at existing for the rest of hypereternity is in jeopardy! So, as the fiction depicts, Parz's worries are reasonable. Similar thinking applies to *Chrononauts*. As the characters stand chatting, they should not expect any changes to the past to hyperpresently affect them. However, they are nevertheless correct to worry—and take action in light of those worries—because of what they hyperexpect to happen.

So, GLOBAL PROPAGATION can help make sense of the metaphysical possibility of certain fictional narratives which we might previously have thought to be impossible.

7. Conclusions

It is interesting to see how exterminous hypertime can (i) avoid problems facing certain other models of time travel and (ii) expand the range of time travel fictions that turn out to be metaphysically possible. I stress again, though, that this is where one's interest should likely stop, for there is no reason to think *our world* is a world of exterminous hypertime. Having said that, were you to ever find yourself in 1930 with a dead Hitler at your feet, then exterminous hypertime should be on your list of candidates as to which metaphysical theory might account for what you have just witnessed.

8. Appendix: The Metaphysics of Time

A referee asked about exterminous hypertime's commitments to the different theories in the metaphysics of time, e.g., must exterminous hypertime be tensed or tenseless, eternalist or presentist, etc.? This appendix deals with those issues, arguing that exterminous hypertime is compatible with any metaphysical theory of time except those committed to an open future.

8.1. Tenseless vs. Tensed Exterminous Hypertime

Consider a tenseless eternalist exterminous hypertime. Given the world is tenseless and eternalist, no time is metaphysically privileged, nor is any hypertime, nor is any time-hypertime[13]. All time-hypertimes exist *simpliciter*, whether they are later, earlier, hyperlater, or hyperearlier. I see no reason to think anything about this tenseless theory is problematic. If anything, it is the most natural understanding of exterminous hypertime.

It gets more interesting when we consider the tensed views. In regular worlds of one-dimensional time, tensed theorists say some time is metaphysically privileged. When we extend that to worlds of exterminous hypertime, it is most natural to think instead that some singular time-hypertime is metaphysically privileged[14].

The rest of this appendix discusses what else we can say about the relationship between exterminous hypertime and the tensed theories of time.

8.2. The Dimensionality of the A-Series

Tensed theorists think that 'privilege forms a series'—namely, the 'A-series'. For instance, at a world of one-dimensional discrete time, consisting of times t_1, t_2, t_3 . . . , then first t_1 would be metaphysically privileged, then t_2, then t_3, etc. Represent that A-series as:

$$t_1 \quad \Rightarrow t_2 \quad \Rightarrow t_3 \quad \Rightarrow \quad \ldots$$

Like the temporal dimension at that world, that A-series is one-dimensional. Tensed exterminous hypertemporal theorists would have a problem if, at worlds of exterminous hypertime, the A-series also had to be one-dimensional. At such worlds, there would be two options as to how such an A-series could be arranged, with each turning out to be a bad option.

Option (i): The series is ordered thus:

$$t_1\text{-}T_1 \Rightarrow t_2\text{-}T_1 \Rightarrow t_3\text{-}T_1 \Rightarrow \ldots \Rightarrow t_1\text{-}T_2 \Rightarrow t_2\text{-}T_2 \Rightarrow t_3\text{-}T_2 \Rightarrow \ldots \Rightarrow t_1\text{-}T_3 \Rightarrow t_2\text{-}T_3 \Rightarrow \ldots$$

This option means that the spotlight of metaphysical privilege must work through every time at the first hypertime before it can move onto the second hypertime and start working through the eternity of times at that hypertime, and so on for all hypertimes. This is problematic because such a series is too similar to the series arising at a world of *one-dimensional time* where, after an eternity has passed, the universe 'resets' to how it was at the start and history starts cycling through again. Imagine that, at such a world, ostensible time travellers who step into their time machines in an earlier circle appear in 'the past' in the next cycle—in a sense, they have not travelled back in time, but have instead travelled an eternity into the future to a point where history has repeated itself. As far as I know, no-one has presented such a theory of time travel, but this short description should suffice in explaining it[15]. Given that theory, the A-series of events would be exactly like that given by option (i). So if the tensed exterminous hypertemporal theorist accepted option (i), I would worry that they were not really distinguishing their theory from this theory of time travel relying only on one-dimensional time.

There is also an apeirophobic problem with the A-series of option (i). Given option (i), metaphysical privilege only moves onto the next hypertime after an eternity has passed. There are Zeno-esque concerns that you will never reach the relevant point where privilege moves to the next hypertime—similarly, it is less clear that you should expect something

to happen if its happening is literally an infinite amount of time away. If that were true then we should never expect metaphysical privilege to move to the next hypertime and so should no longer 'hyperexpect' things that happen in our hyperfuture to come about.

All this said, I assume exterminous hypertemporal theorists will avoid option (i).

Option (ii): The time-hypertimes form the following series, whereby both time and hypertime are always 'flowing forwards':

$$t_1\text{-}T_1 \Rightarrow \quad t_2\text{-}T_2 \Rightarrow \quad t_3\text{-}T_3 \Rightarrow \quad \ldots$$

Given option (ii), some time-hypertimes (e.g., t_1-T_1 or t_2-T_2) are *never* privileged, which seems bizarre. Imagine the standard tensed theorist believed that there were events that are past but which somehow have avoided ever having the light of presentness shine down upon them. That would be crazy! Similarly, it would be weird if, at t_{2021}-T_{2021}, it was true that Hitler was alive in 1930 in virtue of how t_{1930}-T_{2021} is, even though t_{1930}-T_{2021} failed to appear in the A-series and failed to have ever been metaphysically privileged.

So, it is problematic to combine a one-dimensional A-series with two temporal dimensions. However, this is not a problem, since the exterminous hypertemporal theorist should simply say that the A-series is two-dimensional! Given a two-dimensional A-series it makes no sense to think that there is a single time-hypertime that comes 'next' in the A-series. Rather, there is a time-hypertime that comes next in one dimension of the A-series and a different time-hypertime that comes next in the other dimension. For example, if t_1-T_1 is metaphysically privileged then it makes no sense to ask which time-hypertime comes next *simpliciter*. Rather, we should say that t_2-T_1 comes next in one dimension of the A-series whilst t_1-T_2 comes next in another dimension of it. Once we accept that it can be two-dimensional, this problem about ordering the A-series goes away.

Some tensed theorists may struggle with the idea of a two-dimensional A-series. They might, for instance, be wedded to the metaphor of God being sat outside spacetime, in his own (one-dimensional) temporal stream, watching different parts of spacetime 'light up' as the spotlight of metaphysical privilege shines upon it. Such a metaphor leaves no room for a two-dimensional A-series. However, such tensed theorists are surely going to think two-dimensional *time* is also impossible—to think that time can be two-dimensional, but that the A-series must be one-dimensional, seems particularly strange. So, since this paper has already assumed that two-dimensional time is possible, I will assume that a two-dimensional A-series is equally unproblematic.

8.3. Presentist/Moving Spotlight Exterminous Hypertime

Armed with a two-dimensional A-series, exterminous hypertime can be married with different tensed ontologies. Consider a presentist theory whereby only one time-hypertime exists. The non-present non-hyperpresent time-hypertimes do not exist; however—just as other times did and will exist given regular presentism—those time-hypertimes did exist, will exist, hyperdid exist, and/or hyperwill exist. (Additionally, on this view, it makes no sense to ask which time-hypertime will be next *simpliciter* in the A-series, only whether a time-hypertime will be next in in one-dimension of the A-series or the other).

Similar can be said of the moving spotlight theory. Only one time-hypertime is metaphysically privileged. When you are at a metaphysically privileged time-hypertime, the spotlight will next move to different time-hypertimes depending upon whether we consider the dimension of the A-series corresponding to regular time or the dimension corresponding to hypertime—pick a different dimension, and a different time-hypertime will be 'next'.

8.4. Growing Block Theory and the Open Future

Things become more complicated when we consider growing block theory. Firstly, we must pin down the details of what growing block theory amounts to in a world with two temporal dimensions. It is natural for such growing block theorists to say that nothing hyperlater than the metaphysically privileged time-hypertime exists. Equally, it is natural

to say that any earlier-and-hyperearlier, as well as any earlier-and-hypersimultaneous, time-hypertimes exist. However, this does not settle the status of all time-hypertimes. Growing block theorists have two options when it comes to hyperearlier-yet-later time-hypertimes: Either they exist or they do not. Figure 5a depicts a growing block whereby such time-hypertimes *do* exist; call it 'Bigger growing block theory'. Figure 5b depicts a growing block theory whereby those time-hypertimes do not exist; call it 'Littler growing block theory' (because its block is smaller than that of Bigger growing block theory). Both have problems when it comes to their compatibility with exterminous hypertime.

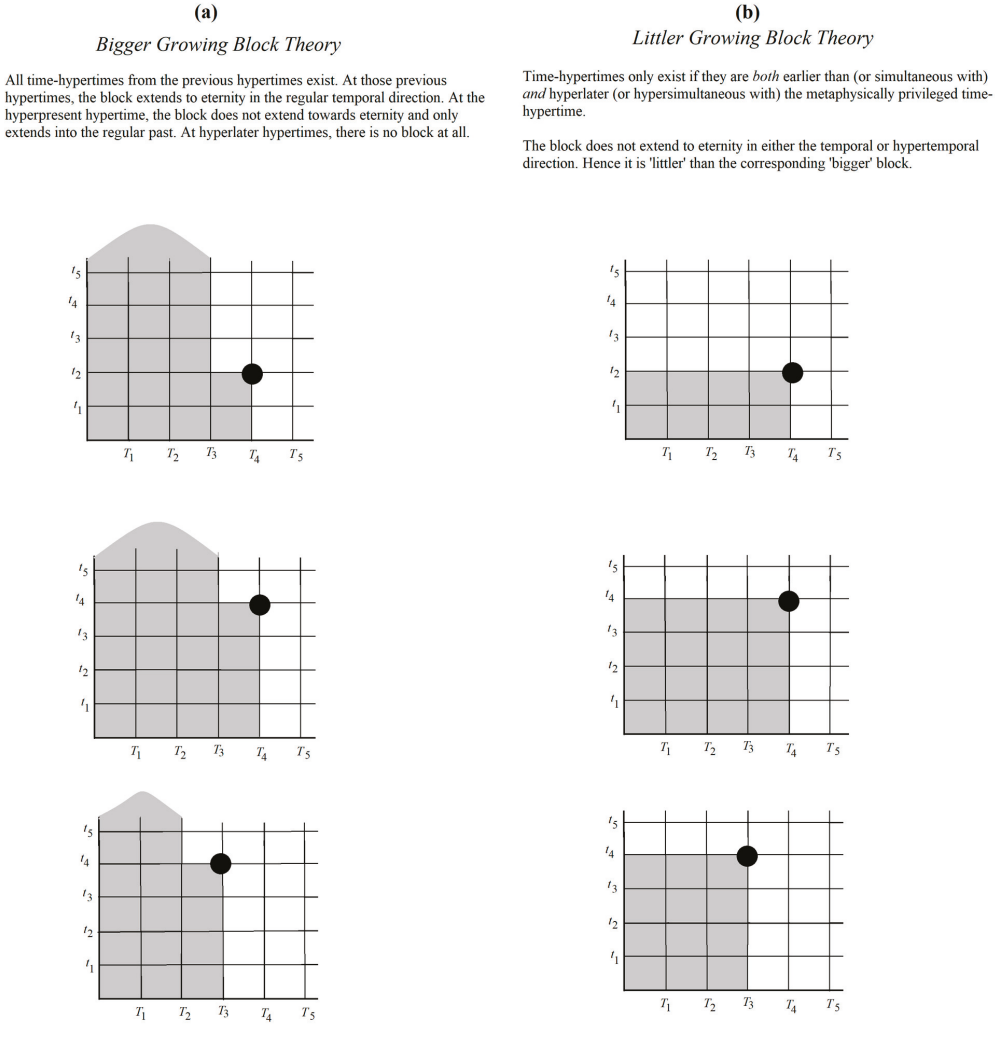


Figure 5. Growing Block Theory & Exterminous Hypertime, (**a**) Bigger Growing Block Theory, (**b**) Littler Growing Block Theory. In each of the diagrams, the black dot indicates which time-hypertime is metaphysically privileged. The greyshading indicates the size of the growing block; every time-hypertime covered by the grey shading exists.

Consider Bigger growing block theory. If all time-hypertimes that are hyperprevious-yet-later exist, then we once again end up with a one-dimensional A-series. To see why,

consider the standard growing block theorist who believes in just one dimension of time. They define the relation of precedence in the A-series as:

> For all j and k: t_j precedes t_k iff the times that exist when t_j is privileged are a subset of the times that exist when t_k is privileged.

Bigger growing block theory can (and, presumably, should!) accept the analogue:

> For all j, k, m, and n: t_m-T_n precedes t_j-T_k iff the time-hypertimes that exist when t_m-T_n is privileged are a subset of the time-hypertimes that exist when t_j-T_k is privileged.

Given that analogue principle, there would then be a one-dimensional A-series like the following:

$$t_1\text{-}T_1 \Rightarrow t_2\text{-}T_1 \Rightarrow t_3\text{-}T_1 \Rightarrow \ldots \Rightarrow t_1\text{-}T_2 \Rightarrow t_2\text{-}T_2 \Rightarrow t_3\text{-}T_2 \Rightarrow \ldots \Rightarrow t_1\text{-}T_3 \Rightarrow t_2\text{-}T_3 \Rightarrow \ldots$$

However, that is just the ordering from option (i) above! Since that ordering was problematic, so too is Bigger growing block theory.

Littler growing block theory also has problems, although this time they stem from the incompatibility of exterminous hypertimes with an open future. Open future theorists believe that facts about later times are indeterminate. Given exterminous hypertime involves two dimensions of time, open future theory must be redescribed. Clearly, that redescription should say that what is hypersimultaneous-and-later is indeterminate, as is anything hyperlater-and-later (and, presumably, anything *at all* that is hyperlater). However, there are two options concerning the status of facts about hyper*earlier*-and-later time-hypertimes:

(a) Facts about hyperearlier-and-later time-hypertimes are determinate; or
(b) Facts about *any* later time-hypertime—whether that time-hypertime is hyperearlier, hypersimultaneous, or hyperlater—are indeterminate.

Option (a) leads to the same problem we had with Bigger growing block theory. Given (a), the facts become determinate in a certain order. All facts at one hypertime go from being indeterminate to being determinate (in order of earliest to latest) and only once facts about all time-hypertimes at any given hypertime are settled, do facts at the hypernext hypertime begin to get settled (again, in order of earliest to latest in the regular temporal series). So there would, again, be a one dimensional A-series of the form:

$$t_1\text{-}T_1 \Rightarrow t_2\text{-}T_1 \Rightarrow t_3\text{-}T_1 \Rightarrow \ldots \Rightarrow t_1\text{-}T_2 \Rightarrow t_2\text{-}T_2 \Rightarrow t_3\text{-}T_2 \Rightarrow \ldots \Rightarrow t_1\text{-}T_3 \Rightarrow t_2\text{-}T_3 \Rightarrow \ldots$$

I have already argued that such a one-dimensional A-series is problematic. So option (a) is a bad option.

Option (b) has its own problem. Consider the following scenario. It is presently and hyperpresently t_{1999}-T_{1999}. Given option (b), facts about hyperearlier/hypersimultaneous earlier times are fixed, so facts about t_{1930}-T_{1998} and t_{1930}-T_{1999} are fixed. Let us say, for example, that it is hyperpresently true that Hitler exists at both those time-hypertimes and at neither does a time traveller turn up to kill him. Given option (b), it is further indeterminate whether, at t_{2021}-T_{1998}, a time traveller leaves that time-hypertime to go and kill Hitler in 1930. Next, imagine time passes and t_{2021}-T_{2021} becomes hyperpresent/present. There would then also be determinate facts about t_{2021}-T_{1998}. Imagine that those facts end up being such that, at t_{2021}-t_{1998}, a time traveller *does* go back to 1930 to kill Hitler. Given PROGRESSION, that time traveller must arrive at t_{1930}-T_{1999}. However, we have assumed for purpose of example that no such time traveller exists at t_{1930}-T_{1999}. So we would now have a contradiction. So option (b) is problematic. Therefore, since option (a) is problematic as well, we cannot pair exterminous hypertime with an open future. (Is it a problem that exterminous hypertime is incompatible with an open future? I presume not. Very few theories of time travel can make room for the future being open [1] (p. 71 n. 3), so it is not concerning that exterminous hypertime ends up in the same boat).

That conclusion in place, return to Littler growing block theory. Stereotypically, growing block theorists are open future theorists—because the future times do not exist, facts about them are unsettled [46] (p. 27–28) [47] (p. 357) [48,49]. Once we add in exterminous hypertime, the natural extension of that stereotype is to say that if a time-hypertime does not exist then facts about it are unsettled. So, if they also indulge in that stereotype, Littler growing block theorists will endorse option (b), which I have argued leads to a contradiction. In short: Unless one is willing to ditch the growing block theorist's traditional commitment to an open future, growing block theory is incompatible with exterminous hypertime. (Similarly, any other tensed theory that allows for an open future will have the same problem—for instance, if you are a presentist or moving spotlight theorist who believes the future is open, then that tensed theory will also be incompatible with exterminous hypertime.)

8.5. Summary

Exterminous hypertime is compatible with the tenseless theory of time and—as long as you are willing to accept a two-dimensional A-series—certain tensed theories of time. The tensed theories that are incompatible are any that accept that the future is open.

Funding: This research received no external funding.

Acknowledgments: Thanks to the two anonymous referees for this journal.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 Even given time travel is physically impossible, some people might think the following question was of interest: 'If time travel were physically possible, then what model of time travel would be true?'. However, that the question is as odd as asking whether, were magic possible, would it work like it does in *Harry Potter* or as it does in *Dungeons and Dragons*? What a bizarre question that is! Similarly, unless time travel *is* physically possible, I doubt any sensible answer to that counterfactual question will be forthcoming.

2 Throughout this paper, I assume PROGRESSION is true. Elsewhere [2], I discuss exterminous hypertemporal worlds at which PROGRESSION is false and 'hypertime travel', where one can travel into the hyperpast, is possible.

3 This footnote discusses the topology of hypertime. A temporal series is discrete iff, with the exception of the first and last instants, every instant has an immediately prior instant and an immediately later instant. (Where x is immediately prior to y iff x is earlier than y and there is no z such that z is earlier than y and later than x; a similar definition applies to 'immediately later'.) A temporal series is dense iff between any two instants there is a third instant [50] (p. 23) [51] (pp. 195–218) [52] (p. 112). Clearly, no dense series is discrete and no discrete series is dense. (Another way of thinking about it is to think of a discrete temporal series as being contiguous to the natural number series, e.g., $\ldots -1, -2, 0, 1, 2 \ldots$, whilst a dense series is contiguous to, e.g., the irrational number series or real number series.) Whilst it might be technically possible to believe in both PROGRESSION and dense hypertime, the most natural interpretation is that time travellers leave one hypertime and move to the 'next' hypertime. Since believing that there is a 'next' hypertime is just to believe that hypertime is discrete, throughout this paper, I assume that hypertime is discrete.

4 Smith intends this to be a problem for Goddu's theory of hypertime [3] (and Meiland's [53], but I have elsewhere suggested that Meiland's theory is not a hypertemporal theory at all [40]). Goddu does not believe hypertime is exterminous, instead believing it to be 'conterminous' (see also [1] (p. 76–77)). Conterminous hypertime is different from exterminous hypertime in that, at any given hypertime, only one regular time exists. Further, as time advances, hypertime advances. Two examples help clarify the difference. Where $\blacktriangleleft$ is the 'hyperearlier than' relation:

Example One: In a time travel case where I kill Hitler, conterminous theorists order the time-hypertimes thus:

$$t_{1930}\text{-}T_{1930} \blacktriangleleft t_{1939}\text{-}T_{1939} \blacktriangleleft t_{2021}\text{-}T_{2021} \blacktriangleleft t_{1930}\text{-}T_{2022} \blacktriangleleft t_{1939}\text{-}T_{2031} \blacktriangleleft \ldots$$

Example Two: In a non-time travel case, e.g., the liberation of Haiti, the regular temporal series is always 'in step' with the hypertemporal series, e.g.,:

$$t_{1803}\text{-}T_{1803} \blacktriangleleft t_{1804}\text{-}T_{1804} \blacktriangleleft t_{1805}\text{-}T_{1805} \blacktriangleleft \ldots\ t_{2021}\text{-}T_{2021} \blacktriangleleft \ldots$$

Given such orderings, Smith's overdetermination worry is not a problem, although for reasons different than for the exterminous theorist. Consider the Haiti example. Exterminous hypertime has a problem because the events at one time-hypertime (i.e., t_{1804}-T_2) are overdetermined in virtue of being caused by events from two distinct time-hypertimes (e.g.,

from the earlier t_{1803}-T_2 as well as the hyperearlier t_{1804}-T_1). No such problem arises given conterminous hypertime. Whilst the conterminous hypertemporal theorist believes that the liberation at t_{1804}-T_{1804} is caused by earlier events *and* caused by hyperearlier events, they are not *distinct* events—the causes earlier in regular time and the causes that are hyperearlier are numerically identical events! Since the events are not distinct, there is not an overdetermination challenge to begin with.

[5] For discussion of the problem of systematic overdetermination being one of unlikelihood, see Funkhouser [54] (pp. 333–338). Other philosophers believe that the problem is something other than unlikelihood (see [55] for discussion); this paper ignores those alternative understandings of the problem of overdetermination.

[6] Smith explicitly likens the problem of overdetermination in the hypertemporal case to the problem of overdetermination in the philosophy of mind (and, when I wrote about the problem [1] (pp. 83–84), I had a similar issue in mind). It is worth noting, then, that when it comes to the problem of overdetermination in the philosophy of mind there are already philosophers who argue against the problem for very similar reasons that I have just given [56,57] (pp. 227–228) [58] (p. 452) [59] (see also [60] and [61] (pp. 722–723)).

[7] A referee raised a worry. Physics says that the half-life of moscovium-287 is 37 milliseconds, i.e., its probability of decaying during that period is 0.5. When I say that its 'objective chance' of decaying in that period is instead 1 and its 'chance$_{\text{chc}}$' of decaying is 0.5, you might think I have become *definitionally confused*. 'Objective chance' picks out just that probability function which physicists are interested in, i.e., the function that says moscovium's probability of decay is 0.5. So what I call 'chance$_{\text{chc}}$' *just is* what we call 'objective chance'; moreover, what I call 'objective chance' must be a totally different function—a function that we might worry is so weird and bizarre that it cannot play a serious philosophical role.The worry is misplaced. Compare to a case where we consider some quantum event which we usually believe has a chance of occurring between 0 and 1. However, imagine it turned out that the 'hidden variable theory' was true, whereby quantum events occur (or not) because of purely deterministic features of the world that we do not have access to. A superscientist who had access to those variables—and who could carry out the appropriate predictions—would see of every event that its chance was either 0 or 1; this would be true, even though more ignorant scientists *justifiably* treated those events as being genuinely stochastic. I claim that there is at least one context/interpretation/understanding whereby: (i) the superscientist is correct to say that the probability function given by the hidden variables is the 'objective chance' function; whilst (ii) the probability function which ignorant scientists are interested in is instrumentally useful, even if it is *not* the objective chance function.If you hold fixed that context/interpretation/understanding, the initial worry of this footnote goes away. A physicist in Case One who was availed of the true laws of nature will, if they suspect they are at a hyperinstant later than T_1, know that all later events have an objective chance of occurring equal to either 0 or 1. However, since there is no time travel in Case One, that physicist will be ignorant of what those chances are. So she—along with her more ignorant colleagues who do not know that later events are nomically enslaved to hyperearlier events—will routinely talk about the moscovium atom having a probability of decaying other than 0 or 1. She correctly recognises that the probability function she is aiming for when she talks in this fashion (i.e., the 'chance$_{\text{chc}}$ function') is not the objective chance function, but it is nevertheless still a perspicuous function that she, and all other scientists, can and should make use of. So I do not think there is any definitional confusion in what I say in the main text.(Additionally, note that there *are* time-hypertimes at which the decay of the moscovium atom is a chancey affair. At t_{2021}-T_1 the objective chance of the atom decaying a few milliseconds later at T_1 is 0.5. That also means that it is *also* true at t_{2021}-T_1 that the atom's objective chance of decaying at t_{2021}-T_2 is 0.5. It is only later on that its chance increases to 1.)

[8] In regular one-dimensional temporal worlds, the chance of an event occurring at an *earlier* time is always 0 or 1, for once the event has/has not happened, it is no longer a matter of chance as to whether it did/did not happen. A similar principle must be true of exterminous hypertemporal worlds—this paper assumes that the chance of events at any hyperearlier time-hypertime is equal to 0 or 1. That is: Even events that occur in your (regular) future are no longer chancey as long as they occurred in the hyperearlier hyperpast.

[9] Footnote 3 argued that hypertime is discrete. Given Global Propagation, regular time must also be discrete because changes to history ripple forward at the rate of one temporal instant per hypertemporal instant, thus one series cannot be continuous whilst the other is discrete. (Discussions of discrete time include [52,62–64] (pp. 114–121); note that, for my purposes, time need only be *possibly* discrete, not *actually* discrete.) One option is that regular time is composed of finitely many 'temporal atoms'. (Such temporal atomicity has been maintained by the likes of: Martinus Capella, the Buddhist Santarankitas, Abu'l–Hasan al Ash'ari, and Abu'l-Mansur al-Maturidi of Samarqand [65]; al-Ghazali and the Mutakallimun, the Greek Epicureans, the medieval philosophers Joannes Canonicus and Nicholas Bonet, and the Jewish philosopher Moses Maimonides [66] (pp. 34–35); (arguably) Descartes [67] (p. 627 n. 2); and the early Russell [68] (p. 6)). Changes to history would then take a finite number of hyperinstants to ripple forwards. For instance, if I time travel to 1930 and kill Hitler then we need only wait until, say, t_{2021}-T_{googol} for 2021 to be such that World War II never occurred. A second option is that time is composed of an infinite, yet discrete, number of instants. In that case, I would have to wait until, say, t_{1930}-T_ω for 2021 to be such that World War II never occurred. (Thanks to Emily Thomas for help with the history of belief in temporal atomicity.)

[10] I trialled a similar solution elswhere [40], but at the time I did not think it worked (for reasons spelt out in that paper). In that paper, I was considering 'past vacillation theory', which is quite different from exterminous hypertime. Nevertheless, the theory of hyper-resilience spelled out in Section 4.2 could probably be tweaked to work for the theory of past vacillation and allow that theory to similarly avoid the Multiple Time Travellers problem.

[11] Following on from *fn*10, I suspect that if changes propagated through vacillating time, rather than exterminous hypertime, we probably *could* allow for the possibility of the narrative of 'Timeslides' (at least, that element of it—there are other elements of the narrative, not discussed in this paper, that might prove problematic).

[12] This raises the question of what it takes for a stage from some future hypertime to be a stage of Nikk Effingham rather than someone else. I have argued elsewhere that there are no great pitfalls to be faced on this issue [1] (see also [5] (pp. 3–4) and [10] (pp. 92–107) for discussion), so I am happy to assume that the hyperfuture stages I pick out are, indeed, stages of Nikk Effingham.

[13] We can rigorously define what a time and a hypertime is:

t is a time $=_{df}$ t is a fusion of the xs whereby: (i) each x is a time-hypertime; (ii) each x is simultaneous with every other x; and (iii) nothing simultaneous with an x fails to be amongst the xs.

T is a hypertime $=_{df}$ T is a fusion of the ys whereby: (i) each y is a time-hypertime; (ii) each y is hypersimultaneous with every other y; and (iii) nothing hypersimultaneous with a y fails to be amongst the ys.

[14] Instead of a time-hypertime being privileged, we might think that a hypertime is privileged (and that *all* time-hypertimes at that hypertime are likewise privileged). A weirder alternative is that the present *time* is metaphysically privileged. I do not discuss these alternatives because they seem substantially less plausible than the view discussed in the main text.

[15] This form of time travel does appear in fiction, e.g., in *Futurama*'s 'The Late Philip J. Fry' [69]. Additionally, whilst I do not know of any metaphysician who proposes such a theory, there is a similar theory in the same ballpark [70] (see also [1] (p. 23)).

References

1. Effingham, N. *Time Travel: Probability and Impossibility*; Oxford University Press: Oxford, UK, 2020.
2. Effingham, N. The Metaphysical Possibility of Time Travel Fictions. *Erkenntnis* **2021**, 1–21. [CrossRef]
3. Goddu, G. Time Travel and Changing the Past (Or How to Kill Yourself and Live to Tell the Tale). *Ratio* **2003**, *16*, 16–32. [CrossRef]
4. Goddu, G. Avoiding or Changing the Past? *Pac. Philos. Q.* **2011**, *92*, 11–17. [CrossRef]
5. Goddu, G. Changing, Annulling, and Otherwising the Past. *Philosophies* **2021**, *6*, 71. [CrossRef]
6. Van Inwagen, P. Changing the Past. *Oxf. Stud. Metaphys.* **2010**, *5*, 3–28.
7. Wasserman, R. *Paradoxes of Time Travel*; Oxford University Press: Oxford, UK, 2018.
8. Smith, N. Why time travellers (still) cannot change the past. *Rev. Port. De Filos.* **2015**, *71*, 677–693. [CrossRef]
9. Hudson, H.; Wasserman, R. Van Inwagen on Time Travel and Changing the Past. *Oxf. Stud. Metaphys.* **2010**, *5*, 41–49.
10. Hudson, H. *The Fall and Hypertime*; Oxford University Press: Oxford, UK, 2014.
11. Lucas, J. A Century of Time. In *The Arguments of Time*; Butterfield, J., Ed.; Oxford University Press: Oxford, UK, 1999.
12. Swinburne, R. *Space and Time*; Macmillan Press: London, UK, 1968.
13. Forbes, G. Review of Oxford Studies in Metaphysics: Volume 5. *Analysis* **2010**, *70*, 571–577. [CrossRef]
14. Baron, S. Back to the Unchanging Past. *Pac. Philos. Q.* **2017**, *98*, 129–147. [CrossRef]
15. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
16. Smith, N. Bananas Enough for Time Travel? *Br. J. Philos. Sci.* **1997**, *48*, 363–389. [CrossRef]
17. Baron, S.; Yi-Cheng, L. Time, and Time Again. *Philos. Q.* **2021**, Forthcoming.
18. Broad, C. *An Examination of McTaggart's Philosophy Vol. II*; Cambridge University Press: Cambridge, UK, 1938.
19. MacBeath, M. Time's Square. In *The Philosophy of Time*; Le Poidevin, R., MacBeath, M., Eds.; Oxford University Press: Oxford, UK, 1993.
20. Richmond, A. Plattner's Arrow: Science and Multi-Dimensional Time. *Ratio* **2002**, *13*, 256–274. [CrossRef]
21. Rutledge, J. Purgatory, Hypertime, & Temporal Experience. *J. Anal. Theol.* **2018**, *6*, 152–161.
22. Schlesinger, G. *Aspects of Time*; Hackett Publishing Company: Cambridge, UK, 1980.
23. Skow, B. *Objective Becoming*; Oxford University Press: Oxford, UK, 2015.
24. Smart, J. The River of Time. *Mind* **1949**, *58*, 483–494. [CrossRef]
25. Tan, P. The Growing Block and What was Once Present. *Erkenntnis* **2020**, Forthcoming. [CrossRef]
26. Thomson, J. Time, Space, and Objects. *Mind* **1965**, *74*, 1–27. [CrossRef]
27. Wilkerson, T. Time and Time Again. *Philosophy* **1973**, *48*, 173–177. [CrossRef]
28. Wilkerson, T. More Time and Time Again. *Philosophy* **1979**, *54*, 110–112. [CrossRef]
29. Bars, I. Survey of Two-Time Physics. *Class. Quantum Gravity* **2001**, *18*, 3113. [CrossRef]
30. Bars, I. Twistor Superstring in Two-Time Physics. *Phys. Rev. D* **2004**, *70*, 104022. [CrossRef]
31. Bars, I. Gravity in Two-Time Physics. *Phys. Rev. D* **2008**, *77*, 125027. [CrossRef]
32. Bars, I.; Deliduman, C. High Spin Gauge Fields and Two-Time Physics. *Phys. Rev. D* **2001**, *64*, 045004. [CrossRef]
33. Bars, I.; Kuo, Y. Supersymmetric Field Theory in Two-Time Physics. *Phys. Rev. D* **2007**, *76*, 105028. [CrossRef]
34. Craig, W.; Weinstein, S. On Determinism and Well-posedness in Multiple Time Dimensions. *Proc. R. Soc. A* **2009**, *465*, 3023–3046. [CrossRef]
35. Vafa, C. Evidence for F-Theory. *Nucl. Phys. B* **1996**, *469*, 403–415. [CrossRef]
36. Effingham, N. An Unwelcome Consequence of the Multiverse Thesis. *Synthese* **2012**, *184*, 375–386. [CrossRef]
37. Colyvan, M.; Garfield, J.; Priest, G. Problems with the Argument from Fine-tuning. *Synthese* **2005**, *145*, 325–338. [CrossRef]
38. Hawthorne, J.; Isaacs, Y. Misapprehensions about the Fine-tuning Argument. *R. Inst. Philos. Suppl.* **2017**, *81*, 133–155. [CrossRef]

39. McGrew, T.; McGrew, L.; Vestrup, E. Probabilities and the Fine-tuning Argument: A Sceptical View. *Mind* **2001**, *110*, 1027–1038. [CrossRef]
40. Effingham, N. Vacillating Time: A Metaphysics for Time Travel and Geachianism. *Synthese* **2021**, Forthcoming.
41. Baxter, S. *Timelike Infinity*; HarperCollins: London, UK, 1992.
42. Millar, M.; Murphy, S. *Chrononauts*; Image Comics: Portland, OR, USA, 2015.
43. Sawyer, R. On the Surface. In *Future Wars*; Greenberg, M., Segriff, L., Eds.; DAW: New York, NY, USA, 2003.
44. Berman, R.; Braga, B. *Star Trek: Enterprise*; Paramount Network Television: Los Angeles, CA, USA, 2001–2005.
45. Naylor, G. *Red Dwarf*; British Broadcasting Corporation (BBC): London, UK, 1988.
46. Tooley, M. *Time, Tense, and Causation*; Oxford University Press: Oxford, UK, 1997.
47. Miller, K. Presentism, Eternalism, and the Growing Block. In *A Companion to the Philosophy of Time*; Dyke, H., Bardon, A., Eds.; Wiley Blackwell: London, UK, 2013.
48. Briggs, R.; Forbes, G. The Real Truth about the Unreal Future. *Oxf. Stud. Metaphys.* **2012**, *7*, 257–301.
49. Longenecker, M. Future Ontology: Indeterminate Existence or Non-existence. *Philosophia* **2020**, *48*, 1493–1500. [CrossRef]
50. Brown, R. *Elements of Modern Topology*; McGaw-Hill: London, UK, 1968.
51. Nerlich, G. *The Shape of Space*, 2nd ed.; Cambridge University Press: Cambridge, UK, 1994.
52. Newton-Smith, W. *The Structure of Time*; Routledge: London, UK, 1980.
53. Meiland, J. A Two-dimensional Passage Model of Time for Time Travel. *Philos. Stud.* **1974**, *26*, 153–173. [CrossRef]
54. Funkhouser, E. Three Varieties of Overdetermination. *Pac. Philos. Q.* **2002**, *83*, 335–351. [CrossRef]
55. Engelhardt, J. What is the Exclusion Problem? *Pac. Philos. Q.* **2015**, *96*, 205–232. [CrossRef]
56. Kroedel, T. Dualist Mental Causation and the Exclusion Problem. *Noûs* **2015**, *49*, 357–375. [CrossRef]
57. Loewer, B. Mental Causation, or Something Near Enough. In *Meaning, Mind, and Matter: Philosophical Essays*; Lepore, E., Loewer, B., Eds.; Oxford University Press: Oxford, UK, 2011.
58. Lowe, E. Substance Dualism: A Non-Cartesian Approach. In *The Waning of Materialism*; Bealer, G., Koons, R.C., Eds.; Oxford University Press: Oxford, UK, 2010.
59. White, B. Metaphysical Necessity Dualism. *Synthese* **2018**, *195*, 1779–1798. [CrossRef]
60. Bernstein, S. Overdetermination Undetermined. *Erkenntnis* **2016**, *81*, 17–40. [CrossRef]
61. Sider, T. What's So Bad About Overdetermination? *Philos. Phenomenol. Res.* **2003**, *67*, 719–726. [CrossRef]
62. Ardourel, V. A Discrete Solution for the Paradox of Achilles and the Tortoise. *Synthese* **2015**, *192*, 2843–2861. [CrossRef]
63. Forrest, P. Is Space-Time Discrete or Continuous? An Empirical Question. *Synthese* **1995**, *103*, 327–354. [CrossRef]
64. Van Bendegem, J. In Defence of Discrete Space and Time. *Log. Et Anal.* **1995**, *150–152*, 127–150.
65. Jammer, M. Concepts of Time in Physics: A Synopsis. *Phys. Perspect.* **2007**, *9*, 266–280. [CrossRef]
66. Turetzky, P. *Time*; Routledge: London, UK, 1998.
67. Levy, K. Is Descartes a Temporal Atomist? *Br. J. Hist. Philos.* **2005**, *13*, 627–674. [CrossRef]
68. Hager, P. Russell and Zeno's Arrow Paradox. *Russell J. Bertrand Russell Stud.* **1987**, *7*, 3–10. [CrossRef]
69. Groening, M. *Futurama*; 20th Century Fox: Los Angeles, CA, USA, 1999–2013.
70. Loss, R. How to Change the Past in One-Dimensional Time. *Pac. Philos. Q.* **2015**, *96*, 1–11. [CrossRef]

philosophies

MDPI

Article

Autoinfanticide Is No Biggie: The Reinstatement Reply to Vihvelin

Richard Mark Hanley

Department of Philosophy, University of Delaware, Newark, DE 19716, USA; hanley@udel.edu

Abstract: David Lewis's attempt to defuse grandfather paradoxes consistently without special restrictions on the ability of time travelers to act in the past is controversial. Kadri Vihvelin uses the case of possible autoinfanticide—killing one's infant self—to argue on Lewisian grounds that Lewis is wrong, since all counterfactual attempts at autoinfanticide would fail. I present a new defense of Lewis against Vihvelin premised on the possibility of personal *reinstatement*, where a person who dies prematurely is replicated from information collected from a previous live scan. I argue on Lewisian grounds that in a Vihvelin case where Suzy does not attempt to kill Baby Suzy, Vihvelin has not shown that Suzy would have failed had she tried to kill Baby Suzy. For, Baby Suzy might have been reinstated. Hence, even granting Vihvelin's own assumptions, a Lewisian can assert that Suzy can kill Baby Suzy. Reinstatement does not require a "big" miracle; so autoinfanticide is no biggie.

Keywords: time travel; reverse causation; fatalism; ability; autoinfanticide; counterfactual dependence; possible worlds; teleportation; personal identity; personal fission; Newcomb Problem

Citation: Hanley, R.M. Autoinfanticide Is No Biggie: The Reinstatement Reply to Vihvelin. *Philosophies* **2021**, *6*, 87. https://doi.org/10.3390/philosophies6040087

Academic Editor: Alasdair Richmond

Received: 15 September 2021
Accepted: 10 October 2021
Published: 18 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

David Lewis [1] argues that even in progenitor or *retro-killing* cases—the most notorious being "grandfather paradox" cases—time travelers have more or less the same abilities as anyone else. In a series of pieces of which [2–4] are representative, Kadri Vihvelin argues that although time traveling to the past and retro-killing is logically possible, in a typical progenitor case the time traveler lacks the ordinary ability to do the deed. In the ordinary sense of "can", a time traveler cannot retro-kill. Ryan Wasserman [5] presents a vigorous recent defense of Vihvelin's view against a range of objections. I take no issue with that defense here, and instead present a new argument to a limited conclusion: that Lewis need not change his own position in response to Vihvelin's arguments. Vihvelin's strategy is to argue against Lewis assuming many of his own views: his account of single timeline time travel, his account of counterfactual dependence, his temporal parts or "worm" theory of persistence, and—though not mentioned explicitly—his account of truth in fiction. But I shall show that Vihvelin has not brought the full suite of Lewisian views to bear on the issue. Considered in the broader light of Lewis's view that teleportation is survivable, that transworld identity is a matter of modal counterpart theory, and that *de re* modality is inconstant, Vihvelin's case is uncompelling, even granting the *Counterfactual Possibility Principle* she proposes to analyze the "can" of ability.

Lewis describes the case of trained assassin Tim who time travels to the past in a single timeline, and who wants his paternal grandfather dead. Can Tim kill Grandfather in 1921, before Tim's father is conceived? Lewis answers Yes, and No. Tim can, in the *ability* sense of "can", kill Grandfather. But Tim will suffer a temporary lack of luck and fail to kill Grandfather, in spite of his ability, because he did after all fail. So in the *luck* sense of "can", Tim cannot kill Grandfather [1]. The crucial assertion Lewis makes is that no systematic explanation of Tim's failure is required—no "boring" temporal censor of the sort others are tempted to invoke [1] (p. 149). Some ordinary occurrence—the world failing to cooperate fully with one's plans—is sufficient. It will be handy to refer to such an occurrence as a

banana peel. Time traveler or not, even the best trained assassin is liable to be foiled by a gun jam or a wind gust or a stray bird or a literal banana peel.

Consider that there are ways for Tim to succeed that are compossible with the killing taking place in 1921, before Tim's father is conceived. Suppose Tim tries to kill Grandfather and succeeds. Unexpected! But it turns out that Grandfather had an arrangement with a sperm bank, and Grandmother was artificially inseminated after 1921. Does ruling out such devices help? Not entirely. Suppose conception happened the old-fashioned way, and by just the man Tim thinks is responsible. Tim nevertheless succeeds in killing him, but it turns out that Grandfather is a time traveler, too. Tim kills him in 1921, before the conception in external time, but not before the conception in Grandfather's personal time. Grandfather had been to the future, visited Grandmother, and ... you get the picture.

The point is that in Lewis's treatment of the grandfather paradox, when he says that Tim cannot kill Grandfather, the progenitor aspect is not after all central to the problem. The real issue is whether or not a single timeline time traveler can change the past—in this case, by causing something to happen on the timeline that never happened on the timeline. That is a contradiction [2]. To eliminate sperm banks and other devices, we need to state the fatalist-sounding view more precisely: Tim cannot kill Grandfather in 1921, *given* that Tim does not kill Grandfather in 1921. But the schema *X cannot do Y at Z given that X does not do Y at Z* is perfectly general, and we should not—on pain of global fatalism—thereby conclude that no one can ever do anything other than what they actually do.

It is no surprise that some philosophers think Lewis has defused the grandfather paradox, while others think he has merely dodged it. Whereas Lewis takes his opponent to be a kind of global fatalist, Vihvelin thinks that Lewis is wrong only about cases sufficiently like retro-killing. Just how far the cases extend beyond retro-killing is an interesting question. Individual human existence seems modally lucky: just about any small change to a range of events preceding your conception would have resulted in your non-existence. Call those events that had to happen just-so for you to exist *fragile*. Where the conception of any of time traveling Tim's progenitors is concerned, not only did Tim not mess with any fragile event, he *could not* have. Thus, Vihvelin's view might have far-reaching consequences for the abilities of time travelers; the fatalism is not global, but quite extensive, going far beyond retro-killings. The stakes are high.

Like Lewis, Vihvelin does not think you will succeed in retro-killing anyone that you did not actually retro-kill. Her distinctive claim is that you could not have succeeded; more precisely, that in a case where you did not try to retro-kill a progenitor on a particular occasion, it is true that had you tried, you would have failed. This raises the further question: what would have stopped you? Her arguments lead Vihvelin to a position that departs significantly from Lewis's. On the one hand, Vihvelin asserts—à la Lewis—that any counterfactual attempt to retro-kill would be foiled by a banana peel; on the other hand, she asserts—*pace* Lewis—that thanks to the nomological impossibility of processes like resurrection, it is the laws of nature that prevent retro-killings.

The structure of this paper is as follows. Section 2 presents Vihvelin's basic argument, which focuses on the case of possible autoinfanticide, and rests upon her Counterfactual Possibility Principle analysis of the "can" of "wide" ability. Section 3 introduces two logical possibilities, personal relocation and personal reinstatement. I show that Lewis believes both are cases of personal survival, and that reinstatement allows for successful autoinfanticide. Section 4 argues that reinstatement is nomologically possible and therefore arguably counterfactually relevant to autoinfanticide cases. Section 5 shows how Vihvelin might reassert the counterfactual irrelevance of reinstatement, by using Lewis's own metric of overall similarity of worlds to argue against the considerations of Section 4. Section 6 rebuts that argument in turn by pointing out that time travelers have counterfactual opportunities to manipulate the past that non time travelers lack. Section 7 shows how to use Lewis's metric to understand his own judgments about counterfactuals in a time travel version of a Newcomb Problem, and applies this understanding to the case of autoinfanticide with reinstatement. I argue that the closest success world is closer by Lewis's metric than any of

Vihvelin's banana-peel failure worlds. Section 8 offers a diagnosis: that the evaluation of these cases is made more difficult by Lewis's own somewhat misleading description of the metric, since "small" miracles need not be small, "big" miracles are not a matter of absolute size, and whether or not a miracle is big or small is highly context dependent. Section 9 uses these considerations to argue that Vihvelin's own position mentioned above—that counterfactual failure to retro-kill would be both effected by a banana peel and forced by the actual laws—is untenable under Lewis's metric. I conclude that Lewisians should continue to say that time travelers can retro-kill in the same sense that non time travelers can kill.

2. Vihvelin's Argument

Can Tim kill Grandfather? Like Lewis, Vihvelin answers Yes and No. Vihvelin allows that it is logically possible to retro-kill. For instance, there are worlds where Tim kills Grandfather and Grandfather is then resurrected [2] (p. 317). So Tim can kill Grandfather. And whereas Lewis says unequivocally Yes, Tim is able to kill Grandfather, Vihvelin answers Yes and No. Tim has the *narrow* ability, in that nothing in Tim's intrinsic properties precludes his killing Grandfather; but lacks the *wide* ability, in that something about Tim's extrinsic properties does preclude his killing Grandfather [3] (pp. 318–319). That makes two senses of "can" for which Vihvelin's answer is Yes, but the dispute is not merely verbal. Vihvelin claims that the wide ability sense is the *ordinary* sense of "can", and moreover that the wide ability sense is not the same as the luck sense Lewis identifies, so that global fatalism does not follow. Non time travelers will often fail to do things in the luck sense, all the while being widely able to do those things, and time travelers will often fail to do things in the luck sense, all the while being widely able to do those things. But time travelers have the distinction of sometimes failing to do a thing because they lack the wide ability to do it, *because* they are time travelers [3].

To distinguish between the luck or logical possibility senses and the wide ability sense, Vihvelin employs a principle not found in Lewis's work. Vihvelin's most comprehensive statement of it is as follows [3] (p. 319):

> *Counterfactual Possibility Principle*: S has, at time *t*, the wide ability to A only if it's *not* true, at t, that if S tried (again) to A, S would fail.
>
> [emphasis original]

This is equivalent to saying that S can do A only if were S to try (again) to A, S might succeed. The point of "again" is to allow for temporary lack of luck. If Tim were indeed widely able to kill Grandfather, then if Tim at first failed to shoot Grandfather dead because Grandfather's cigarette case deflected the bullet meant for his heart, it would be true that were Tim to try again, he might succeed. But, Vihvelin thinks, Tim would fail no matter how many times he tried. So Lewis [1] (p. 150) is wrong to say that Tim's failure even once is due to temporary lack of luck rather than any lack of ability.

The Principle also distinguishes between wide ability and logical possibility. By Lewis's own analysis, a counterfactual (or more generally, a subjunctive conditional) of the form "If it were that *p*, then it would have been that *q*" is actually true if and only if a world where *p* is true and *q* is true is, on balance, closer to the actual world than any world where *p* is true and *q* is not true. Vihvelin applies this schema to the example of autofanticide, which cannot be dodged by introducing a sperm bank or making the progenitor a time traveler. She imagines adult Suzy, who is sent back in time to visit her infant self and fails five times to kill Baby Suzy. Would Suzy have succeeded in a sixth attempt? No. Like Tim, Suzy would always have failed, no matter how often she tried. Therefore, by the Counterfactual Possibility Principle, Suzy is unable to kill Baby Suzy. Why would she always fail? Vihvelin writes [3] (p. 322):

> The worlds where Suzy tries and succeeds are worlds which either have resurrection from the dead or some sort of system of ontological understudies.

Call any "ontological understudy" case a *replacement*, where the identity between Suzy and Baby Suzy is broken in the world where Suzy's attempt succeeds. Vihvelin then gives an argument by cases. A world in which Suzy succeeds and in which Baby Suzy is resurrected is logically possible but nomologically impossible, and such a world is too distant to underpin counterfactual success. A replacement world is either not a world in which *Suzy* kills *Baby Suzy*, and so is utterly irrelevant, or if it somehow does count as a success world it involves breaking the causal dependence of adult Suzy on Baby Suzy, and hence such a world is too distant to underpin counterfactual success.

I am going to simplify things a little. Focus on what I shall call *Good Suzy World*, where 30-year-old time traveler Suzy is very healthy and strong and quite reasonably makes no attempt at all on Baby Suzy's life. She visits Baby Suzy and is alone with her for 10 min (their parents are downstairs). Baby Suzy is asleep in an open crib, and having locked the door and window, Suzy leans in at time t and gently, silently chucks Baby Suzy under the chin. No banana peels are present to get in Suzy's way. While Baby Suzy continues sleeping, Suzy unlocks things and leaves, and then time travels back to the future, never to return. If Vihvelin is right about Good Suzy World, Suzy not only does not but cannot kill Baby Suzy at t. Assume there is one closest possible world where Suzy tries to kill Baby Suzy by crushing her windpipe instead of chucking her under the chin; call this *Bad Suzy World*. According to Vihvelin, it is a world where Suzy's attempt fails.

3. A Lewisian Counter: Personal Reinstatement

A series of metaphysical objections to Vihvelin have defended Lewis by appeal to cases of replacement, some of them quite exotic [4]. For instance, Peter Vranas argues that Vivhelin's argument is refuted by attending to a relevant metaphysical possibility [7] (pp. 118–119):

> [C]onsider a world—to simplify, and without loss of generality, say it is the actual world—at which Baby Suzy has an identical twin, Twin Baby Suzy, and at which Suzy sets off a bomb in a room where Baby Suzy and Twin Baby Suzy are asleep, intending to kill them both, but the bomb happens to kill only Twin Baby Suzy. Consider also a world w which is qualitatively identical to the actual world, but at which (i) the bomb happens to kill only Baby Suzy, and (ii) Suzy is a later stage of Twin Baby Suzy, not of Baby Suzy . . . . Then w is a world at which Suzy tries to kill Baby Suzy and succeeds, and at which Suzy is a later stage of some baby-stage (namely Twin Baby Suzy) whose DNA matches the DNA of Baby Suzy not by some miracle or improbable coincidence, but rather because the two baby-stages are identical twins. Since w is qualitatively identical to the actual world, w is at least as close to the actual world as any world at which Suzy tries to kill Baby Suzy but fails.

This is a difficult case to understand. It seems to appeal to *haecceitism*; and, if so, Lewis would reject it [8] (pp. 220–235). Lewis allows that you can coherently contemplate the possibility of being someone else, exactly as they actually are, but that is not contemplating some distinct possible world [8] (pp. 231–232); and it is not a possibility where you are both you and them. But let us suppose that Lewis can be persuaded that w is a possible world distinct from the actual world. It is still not clear that this possibility would refute Vihvelin. For, w would have to be as close to the actual world as the actual world is, since the nearest world where Suzy tries to kill Baby Suzy but fails is by hypothesis the actual world. Perhaps a difference in who is who does not matter to how close a world is to the actual world when assessing subjunctive conditionals, and Vranas can truly say that Suzy *might* have succeeded. (He cannot say she *would* have). But, on balance, I think Lewisians should prefer a more convincing rebuttal.

Fortunately, there is for Lewis a much more promising metaphysical possibility. Vihvelin allows that there are resurrection worlds where Suzy succeeds: they are worlds at which Baby Suzy dies and is buried, but is later resurrected from the dead and grows up to be the adult Suzy who travels back through time and kills her baby self [2] (p. 321).

Note what would not count as success. Had Suzy tried to drown Baby Suzy by throwing her into a frozen lake, it might be a case where the heart stops through hypothermia, and yet the victim can be revived by (carefully) warming them up again. In such a case Baby Suzy has not actually died, so this is not a success world. Nor is a world where Suzy tried to kill Baby Suzy by placing her in suspended animation, with Baby Suzy subsequently being reanimated. Nor is any *Princess Bride* world, where people can be "only mostly" dead. Vihvelin is instead thinking of more drastic, even colder cases where the body's individual cells have died and have decomposed and are somehow brought back to life through a reversal of that decomposition. Call this *corpse-resurrection*. I grant Vihvelin's claim that corpse-resurrection worlds are nomologically impossible and too distant for Suzy's counterfactual success.

But imagine a different kind of world: a teleportation world or *T-world*, where humans employ scanning and replication technology, in the first instance to travel. Scanning at the departure point is instantaneously followed by deliberate, total bodily destruction—which surely does count as death—and then single replication at the destination. Call this *relocation*. Is relocation resurrection of the same person, or is it merely replacement by a doppelgänger? For Lewis, it is resurrection. He writes concerning the question of what matters in personal survival [9] (p. 17):

> I answer, along with many others: what matters in survival is mental continuity and connectedness . . . . My total present mental state should be but one momentary stage in a continuing succession of mental states. These successive states should be interconnected in two ways. First, by bonds of similarity. Change should be gradual rather than sudden . . . Second, by bonds of lawful causal dependence . . . . [E]ach succeeding mental state causally depends for its character on the states immediately before it.

Lewis believes in person stages which are proper temporal parts of persons. When the two kinds of bond are present between two stages, they are R-related. In relocation a T-world traveler is temporally gappy, but if the first stage of the replica is R-related to the last stage of the scanned subject, then the gap is no obstacle to survival, and the traveler is one person. Lewis explicitly endorses teleportation as survival [10] (pp. 192–193):

> Consider our opinions about teletransportation, an imaginary process that works as follows: the scanner here will take apart one's brain and body, while recording the exact state of all one's cells. It will then transmit this information by radio. Traveling at the speed of light, the message will reach the replicator. This will then build, out of new matter, a brain and body exactly like the one that was scanned. Some philosophical positions on personal identity imply that one survives teletransportation (unless it malfunctions). Others imply that teletransportation is certain death. Now, imagine that a philosopher is caught on the seventeenth story of a burning building. He has some hope, but no certainty, of the ordinary sort of rescue. Then he is offered escape by teletransportation, provided he accepts the invitation right away. At that point, I think his philosophical opinion may very well guide his decision. *If he thinks what I do, he will accept teletransportation* even if he reckons his chance of ordinary rescue to be quite high.
>
> [footnotes omitted, emphasis added]

Suppose that in the T-world humans are also regularly scanned as insurance against unforeseen death (e.g., by murder or bad accident); call any consequent replication *reinstatement*. The apparent difference between relocation and reinstatement is that in the latter, the scanned subject has two continuers: the short-lived stage whose death prompts the replication; and the longer-lived replica. Is reinstatement resurrection or replacement? To begin, here is Lewis in his final publication [11] (p. 12):

> Suppose you are about to be beamed up, and you know that the signal will be received both on the starship *Enterprise* and on the starship *Potemkin*. Let's assume that beaming up works not by transmission of matter, but by transmission

> of structural information. That guarantees causal continuity in all bodily and mental respects. You will survive twice over. (What does it matter that you will be made of different atoms afterward? Atoms are the ultimate interchangeable parts, and most of them will be replaced within a few years anyway). Should you expect to find yourself aboard the *Enterprise* or aboard the *Potemkin*? Both. One of your future selves will be aboard one and another will be aboard the other . . . .
>
> Suppose you're about to be beamed up, with the signal received both on the *Potemkin* and on the *Enterprise*. At the last moment you find out that the receiver on the *Enterprise* is malfunctioning: anyone transported there will be dead on arrival, or very soon after. What to expect? No worries, you'll be safe and sound aboard the *Potemkin*. Your death branch should not figure in your expectations.

Here Lewis seems to be drawing upon his treatment of fission in [9,12] where he argues that the R-relation is near enough to identity. The I-relation holds between two temporal parts of one and the same person, and Lewis argues that the R-relation is the I-relation. Since strict identity and its cognate I-relation are each symmetric, Lewis posits a symmetric R-relation. Lewis writes [9] (pp. 23–24):

> If a stage S_2 is mentally connected to a previous stage S_1, S_1 is available in [quasi-] memory to S_2, and S_2 is under the [quasi-] intentional control of S_1 to some extent—not the other way around. We can say that S_1 is R-related *forward* to S_2, whereas S_2 is R-related *backward* to S_1 . . . . S_1 and S_2 are R-related *simpliciter* if and only if S_1 is R-related either forward or backward to S_2 . . . .
>
> In a case of fission, for instance, we have a prefission stage that is R-related forward to two different, simultaneous postfission stages that are not R-related either forward or backward to each other.
>
> [emphasis original]

Hence, the R-relation is not transitive. To illustrate this, suppose that Yuri is beamed from the *Hood* and is the one who knows that he will have two long-lived continuers on the *Enterprise* and *Potemkin*. Yeva is beamed from the *Hood* and knows that she will have one long-lived continuer on the *Potemkin* and that her replica on the *Enterprise* will arrive alive and awake but die soon thereafter [5]. In both cases, suppose that Yuri, Yeva and their continuers never undergo any other fissions or fusions. Since a person is a maximal aggregate of R-related person-stages, Lewis would say that the Yuri and Yeva cases are fissions each involving exactly two persons. In each case, a person-part exists up until the scan: call them Yuri_H and Yeva_H. After the replications, in each case, there are two distinct stages on board each ship; call them Yuri_E and Yuri_P, who are not R-related to each other, and Yeva_E and Yeva_P, likewise not R-related to each other. The Yuri case has two persons, (Yuri_H + Yuri_E) and (Yuri_H + Yuri_P); the Yeva case also has two persons, (Yeva_H + Yeva_E) and (Yeva_H + Yeva_P). On Lewis's view, the pre-fission Yuri_H was a common part shared by two persons; ditto for Yeva_H. Each person sharing Yuri_H wants to survive beaming, and their desire to survive must include a plural desire; of the *strong* form *let all of us survive* or of the *weak* form *let at least one of us survive*. In the Yuri case, both the strong and the weak forms would be satisfied. But Yeva is a case of survival, too. According to Lewis, there is a weak ordinary desire to survive that is satisfied in both cases [12] (pp. 75–76).

Back on our T-world, suppose that Stan has never relocated, but is scanned regularly. At the age of 25, Stan is murdered, dying instantly exactly one day after his most recent scan. That scan is used to reinstate Stan exactly one day after the murder. Call the 25-year-old worm that exists up until the scan $Stan_1$, the one-day stage between the scan and the murder $Stan_2$, and the replica $Stan_3$; and for simplicity suppose that Stan_3 is never relocated or reinstated. The Stan case seems more like the Yeva case than the Yuri case, so it seems we should interpret it as a fission case with a death branch. That seems to be Lewis's view when discussing an analogous case [12] (p. 75):

> [C]onsider a system of survival insurance . . . . From time to time your mind is recorded; should a fatal accident befall you, the latest recording is played back into the blank brain of a fresh body . . . . [T]he fission occurs at the time of recording . . . . This system satisfies the weak desire for survival, but not the strong desire.

Aggregating R-related stages we count two overlapping persons. $Stan_1$ wants to survive, but as a shared stage has the weak plural desire *let* ($Stan_1$ + $Stan_2$) *or* ($Stan_1$ + $Stan_3$) *survive*. By day two after the murder, only the latter survives, but that is good enough for Lewis. With this in place, suppose that Good Suzy World is a T-world, and that like Stan, Suzy is not a relocator. Suzy does not try to kill Baby Suzy, and Baby Suzy was regularly scanned but never reinstated. If Suzy had tried and succeeded in killing Baby Suzy, then no biggie: like Stan, Baby Suzy would have been reinstated from the most recent scan exactly one day before *t* (the time of death), and the replica who began exactly one day after *t* would have grown up into (Bad) Suzy. So Suzy can kill Baby Suzy.

But there is a problem. Stipulate that Baby Suzy already counts as a person, and already has whatever counts as the ordinary desire to survive. In Good Suzy World, Baby Suzy is one and the same person as Suzy, so that Suzy killing Baby Suzy would have to count as autoinfanticide. Bad Suzy in Bad Suzy World just described is closely analogous to Stan in Good Suzy World; but we might be troubled by her relevance to the *Good Suzy* story. By hypothesis, Good Suzy has no shared stages, but she nevertheless has the weak plural desire for survival (since according to Lewis that is a desire that ordinary folk have). In Bad Suzy World, there are two persons sharing the Baby Suzy stage, BS_1. There is the one-day stage BS_2, and there is the replica BS_3. On the fission reading, the two persons are (BS_1 + BS_2) and (BS_1 + BS_3). It seems plausible by Lewis's account that adult Bad Suzy is adult Suzy. But who kills who? BS_3 kills BS_2, and they are (parts of) two different persons! Hence, on the fission hypothesis, Bad Suzy World seems not to be a world in which Suzy kills Baby Suzy—indeed, it is not an autoinfanticide world at all. It is a replacement world.

Two responses are open to Lewis. The first is to appeal as Lewis does to a general inconstancy in *de re* modal judgments— such as saying *this* thing could have killed *that* thing—claiming that the same actual thing can be multiply represented at another possible world. For Lewis, cross-world judgments of personal identity are analyzed in terms of a modal counterpart relation mediated by resemblance, and the counterpart relation does not always behave like the identity relation. Which way we represent *de re* is heavily affected by context, governed by a Rule of Accommodation according to which there is a presumption that utterances be interpreted as true [8] (pp. 248–263). But whatever success this inconstancy reply enjoys, as long as we interpret the Bad Suzy case as a fission, it remains true that Bad Suzy does not commit autoinfanticide. Hence, I shall pursue a different strategy.

Reinstatement Restated

The strategy is to argue that Lewis can maintain the relevance of Bad Suzy World without invoking inconstancy, by denying that the Bad Suzy case is a fission. It is time to reconsider how we view a reinstatement like Stan's. Lewis has in effect described four cases: the Yuri case with two life branches; the Yeva case with one life branch and one death branch; the relocation case; and (in a slightly different version) the Stan case. The Yuri case is definitely a fission; the relocation case is definitely a nonfission. Lewis does not explicitly say that the Yeva case is a fission, but we should presume he thinks so since he explicitly says that a case like Stan's is a fission.

In discussing fissions Lewis is mainly concerned with the forward-looking attitudes of the prefission subjects. Consider instead some backwards-looking attitudes. Suppose that Stan at 50 has two life-long friends. Dan, who is also 50, has never relocated or been reinstated. Dan contemplates what would have happened had he been killed at 25. He knows the facts about the timing of his scans, and judges that had *he* been killed at a certain moment at 25, *he* would have been reinstated from the most recent scan, taken 24 h earlier.

Is he wrong? After all, Dan reasons, that is what happened to Stan: *he* died at 25 and *he* was reinstated. On the other hand, their friend Ivan had not been scanned since he was 25, but had no need of relocation or reinstatement until just last week, when killed at the age of 50. 24 hours later, a replica was produced from the 25-year-old scan, and that replica is now chatting with Dan and Stan. The replica says that *he* was killed last week, and *he* was reinstated. Is he wrong? Yes, Dan reasons. The replica looks 25 and does not even remember the last 25 years of their friendship. Whoever he is, *he* was not killed last week. Dan remembers how grateful he was to get his friend Stan back after reinstatement. But this replica is no Ivan. In fact, he's only making things worse.

I think a Lewisian can make sense of Dan's judgments. The Stan case is quite unlike the Yuri case. It shares with the Yeva case the existence of one life branch and one death branch, but the death branch does not co-exist at the same external time as the life branch. In that respect the Stan case is more like relocation. Moreover, there is no causal dependence of the life of Yeva_{H} on the death of Yeva_{E}. By contrast, Stan_3 only exists because Stan_2 dies. The last stage of Stan_2 and the later first stage of Stan_3 are very, very similar. The similarity is no coincidence, and very strong bonds of counterfactual dependence are in place: had Stan_2 been mentally very different then so would Stan_3 have been mentally very different. So the death of Stan_2 causes the existence of a stage whose mental states are counterfactually highly dependent upon Stan_2's mental states, and the counterfactual dependence is normal in direction, with states later in external time counterfactually dependent upon earlier states. The Lewisian instinct is after all to analyze causal dependence in terms of counterfactual dependence [13]; so for Lewis, Stan's reinstatement is quite close to relocation, and quite close to ordinary survival.

The Ivan case is different again. Call the worm that lives until 25 Ivan_1, the worm from 25 to 50 Ivan_2, and the replica Ivan_3. Ivan_2 does not coexist at the same external time as Ivan_3, so his case is in that respect unlike the Yuri and Yeva cases, and more like the Stan case. It is also like the case of Stan in that the death of Ivan_2 causes the existence of Ivan_3. But the Ivan case is quite unlike the Stan case in that the bonds of similarity and counterfactual dependence between Ivan_2 and Ivan_3 are very much weaker. The lesson seems to be that very short-lived death branches are unproblematic, but the longer they last, the more problematic they become [6].

So I believe that the Lewisian can say that reinstated Stan is one person consisting of Stan_1, Stan_2, and Stan_3, with a one-day gap in his existence. And if that is true of Stan, then Bad Suzy is a single person consisting of BS_1, BS_2, and BS_3, and Bad Suzy World is straightforwardly a world where Bad Suzy commits autoinfanticide. So *she* died and *she* was reinstated. To summarize, I have argued that there are two potential ways for Lewis to endorse reinstatement as a means of personal survival, and to hold that there are possible worlds where Suzy kills Baby Suzy and Baby Suzy is reinstated. Now, I must make such worlds relevant to the assessment of Vihvelin's counterfactuals.

4. The Relative Closeness of Reinstatement

Suppose that Vihvelin grants that relocation or reinstatement at a T-world is survival. She might yet claim that T-worlds are still not relevant to the counterfactuals, even for Lewis, since Good Suzy World is not a T-world. In one version, this response might assert that relocation and reinstatement though logically possible are nomologically impossible. But Lewis is not bound to agree. Reinstatement is at most *technologically* impossible, and even that is doubtful. We already know how to kill, so we just need advanced enough scanners, and advanced enough 3-D printers [7]. Whereas Vihvelin thinks that time travel worlds are more like ours than any resurrection worlds are, for Lewis that judgment if anything seems to be reversed. Vihvelin [2] (p. 323) writes "I think that time travel is possible at worlds very much like ours, maybe exactly like ours"; whereas according to Lewis [1] (p. 145), "a possible world where time travel took place would be a most strange world, different in fundamental ways from the world we think is ours" [8].

Suppose Vihvelin grants that relocation and reinstatement are nomologically possible. And note two things. Good Suzy World is described in a fictional story—*Good Suzy*, told above—and according to Lewis's own account [14], what is true in a fictional story is what would have been true had the story been instead told as known fact [9]. Applying Lewis's own treatment of counterfactuals to that analysis, if the actual world is not a T-world, then Good Suzy World is not a T-world either, since adding relocation or reinstatement would be a gratuitous change. Moreover, if Good Suzy World is not a T-world, then neither is Bad Suzy World. Hence, Bad Suzy would fail to kill Baby Suzy.

If Good Suzy World is not a T-world, then neither is Bad Suzy World; merely attempting to kill someone will not license adding relocation or reinstatement technology to the world. But the other two steps in the argument just given on behalf of Vihvelin are dubious. Lewis does not think that time travel to the past occurs in the actual world, but he might think that our world is a T-world simply awaiting more advanced technology. And if so, then Good Suzy world—where time travel technology has been developed—likely is a T-world, too. But suppose ours is not a T-world. Vihvelin claims [2] (p. 323):

> But in any case there is no reason to suppose that there is any connection between time travel-permitting laws and resurrection-permitting laws.

This is plausible for the corpse-resurrection Vihvelin has in mind, but less so for relocation or reinstatement. Lewis leaves room for cases of *instantaneous* time travel, where a journey through external time takes no personal time at all [1] (p. 146). Lewis does not describe any such cases. He describes only devilish *Fred-cum-Sam* cases, which might appear to be time travel but are not [1] (p. 148). Consider instead DerfMas: Mas is born and lives a normal life until he is vaporized by a time-traveling demon who remembers his entire final qualitative state, and then uses that knowledge to produce Derf in the past. Derf appears as if in the midst of life, and lives normally from then until an ordinary death, before Mas is born. The demon ensures that Derf's initial qualitative state exactly (or as much as is physically possible) resembles Mas's final state. DerfMas is one person, and time travels instantaneously to the past, by Lewis's account.

Derfmas is nomologically impossible, thanks to the demon. But as long as the laws permit information to be sent into the past, there is no need for the supernatural; relocation can be used to send time travelers to the past. If ours is not a T-world, and granting Lewis that it is not a time travel world either, then Good Suzy World would be a T-world if the closest time travel worlds employ instantaneous relocation time travel. Nothing in Vihvelin's argument rules this out. I conclude that Vihvelin so far fails to show that Bad Suzy *would* fail to kill Baby Suzy.

5. The Metric of Overall Similarity of Worlds

Vihvelin [4] employs Lewis's own account [15] of the metric of overall similarity of worlds designed to be plugged into his "Analysis 2" of counterfactuals. Lewis supposes—as I shall in what follows—that the laws of nature are deterministic, but that nomological possibility nevertheless allows that on many occasions things could have happened differently. Pick one such occasion, at time o. Lewis describes four different kinds of possible world where the antecedent of a standard counterfactual about what happens at o is true (in each case, I will assume that the centered world is our actual world). The one that counts as closest, w_1—I will call it a *first-rate* world—is in external time exactly like the actual world up until just before o, when there is one "small miracle"—a departure from our laws—and divergent thereafter (divergent with respect to the pattern of events, not with respect to the laws). A *second-rate* world w_2 matches our deterministic laws exactly, but differs in what happens at o, and so is never exactly like ours in matters of particular fact. A *third-rate* world w_3 is exactly like ours until just before o, when there is a small miracle, and then approximately but not exactly like ours thereafter, in virtue of a second small miracle that prevents the more drastic consequences of the first small miracle. A *fourth-rate* world w_4 is exactly like ours until just before o, when there is a small divergence

miracle, followed immediately by a big convergence miracle which in effect undoes the small miracle, and so is exactly like ours again thereafter. Lewis writes [15] (p. 472):

> Under the similarity relation we seek, w_1 must count as closer to [the centered world] than any of w_2, w_3, and w_4. That means that a similarity relation that combines with Analysis 2 to give the correct truth conditions for counterfactuals such as the one we have considered, taken under the standard resolution of vagueness, must be governed by the following system of weights or priorities.

(1) It is of the first importance to avoid big, widespread, diverse violations of law.
(2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(3) It is of the third importance to avoid even small, localized, simple violations of law.
(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

At step (1) we eliminate fourth-rate worlds, at step (2) we eliminate second-rate worlds, and at step (3) we eliminate third-rate worlds, leaving a first-rate world as the closest, on balance, to the actual world. It is crucial to Lewis's account that we balance both similarity in the pattern of events *and* similarity in the laws. This is to block the "future similarity objection" that a metric of overall similarity must favor third-rate worlds where a second small miracle prevents a more drastic future difference in the pattern of events. Anticipating a little, call such third-rate worlds *banana peel* worlds. Grant that if the centered world is the actual world, had Nixon pressed the wrong button, then there would have been a nuclear holocaust—the system was reliably set up that way—and grant also that there never will be an actual nuclear holocaust. In a banana peel world, Nixon presses the button, but a banana peel that does not actually exist *would have* existed to prevent the holocaust. If we count only the pattern of events, then given the difference that a nuclear holocaust would make, a banana peel world arguably comes out closest, such that the counterfactual we began with wrongly comes out false. However, once we include the laws in the metric, then given the difference to the laws that a second small miracle makes, banana peel worlds lose, given condition (3) and buttressed by condition (4).

Vihvelin [4] summarizes her employment of Lewis's metric as follows (emphases original):

> The closest worlds where Suzy's attempt to kill the baby succeeds are worlds with *one small* and *one big miracle*, whereas the closest worlds where Suzy's attempt fails are worlds with, at most, two *small miracles*. Since Lewis's theory says that worlds with one or even two small miracles are closer than worlds with one big miracle, his theory says that worlds where Suzy's attempt fails are closer than worlds where her attempt succeeds [10].

Well, hold on. Vihvelin is right that there are possible worlds where a baddish Suzy tries to kill Baby Suzy and fails with no need of a second small miracle. But such worlds are first-rate only with respect to centered worlds that have built-in fail-safes—booby traps, motion-detectors, and the like—that would have stopped almost anyone from killing Baby Suzy. That includes an intrinsic duplicate of Suzy; yet Vihvelin grants that such a duplicate in Suzy's place could have killed Baby Suzy [2] (p. 327). Hence, such fail-safe worlds are irrelevant to Vihvelin's argument, since the centered goodish Suzy world where we evaluate the counterfactuals will also have those fail-safes. By contrast, I stipulated that Good Suzy World has no fail-safes, so Bad Suzy World has none, either. Hence, Vihvelin is committed to Bad Suzy failing because of a banana peel that does not exist at Good Suzy World.

On the other side of the ledger, Vihvelin need not deny the nomological possibility of relocation or reinstatement. Good Suzy World does not serve Vihvelin's argument if it is a T-world such that Baby Suzy has been recently scanned. That gets Lewis a first-rate world too easily—his own built-in fail-safe—and Suzy can kill Baby Suzy. This seems to at least narrow the scope of Vihvelin's conclusion, but to be fair, let's make things as bad as we can for Lewis. Suppose that Good Suzy World is a T-world, but stipulate that Baby

Suzy's parents are fanatical members of the van Inwagen Society and have refused to ever have Baby Suzy scanned. Now, there is trouble, for Vihvelin will claim that any success worlds will have to be big miracle worlds. (They will not be quite fourth-rate, since they do not require the entire world to reconverge on the centered world, but they are very, very distant from the centered world). Vihvelin would claim in such a case that Bad Suzy World is therefore a world in which Suzy fails. But what stops Bad Suzy? A second small miracle. So Bad Suzy World is a third-rate, banana peel world.

Or is it? A gap remains in Vihvelin's argument, for there is another important class of worlds it has not yet taken into account.

6. Bad-but-Smart Suzy

In the movie *Bill and Ted's Excellent Adventure*, the eponymous teens when faced with an uncooperative world repeatedly take advantage of the fact that they can time travel to produce desired outcomes. At one point they need Ted's father's keys to the police station, so they decide at time k to later use their time machine to travel back to two days before k and borrow the keys, which they can then leave behind a nearby sign to be found at k. They look behind the sign and retrieve the keys. The world must after all cooperate, and part of the setup is that the keys in fact have at k been missing for two days. And of course Bill and Ted or someone else must follow through with the future time travel journey, and get the job done in the past.

Consider a possible world where *Smart Suzy* behaves like Bill and Ted. Smart Suzy World follows the *Good Suzy* story up until she chucks Baby Suzy under the chin, but then a villain breaks into the room, kills Baby Suzy, and escapes. Unexpected! As far as Smart Suzy knows her parents never had her scanned; but she must have been scanned, anyway. Someone must have arranged it, but who, and why? Smart Suzy leaves discreetly and visits a nearby scan bank. Sure enough, they have a recent scan of Baby Suzy on file and Suzy orders reinstatement. Suzy then uses her time machine to travel to a time earlier than her first visit, briefly kidnaps Baby Suzy and gets her scanned. (The order of these events in Suzy's personal time could be altered: she can go back to get the scan made before she orders the replica made). Now, apply this reasoning to the counterfactuals true in the *Good Suzy* story. Big and strong as she is, Bad Suzy crushes Baby Suzy's windpipe; her attempt to retro-kill succeeds. But Bad Suzy is also smart, and thereafter she uses time travel and reinstatement to ensure her own survival. She sneaks the replica back into the crib, thereby relieving her none-the-wiser parents of their grief, and leaving them believing that something like corpse-resurrection has occurred [11].

The question is how a Bad-but-smart Suzy World fits into Lewis's metric when it is centered on Good Suzy World. I shall argue that a Bad-but-Smart Suzy World is first-rate, and so beats out any banana peel world. So it is true in *Good Suzy* that had Suzy tried to kill Baby Suzy, someone would have had to have time traveled to a previous time and arranged a scan for Baby Suzy. (I am here assuming a competent attempt). Even if that is false, had Suzy tried to kill Baby Suzy, she *might* have succeeded, and so it might have been that someone would have had to have time traveled to a previous time and arranged a scan for Baby Suzy. Given the Counterfactual Possibility Principle, by Lewis's own account, it is true in *Good Suzy* that Suzy can kill Baby Suzy. No big miracle needed; autoinfanticide is no biggie.

Or so say I. Lewis never tells us explicitly how time travel counterfactuals are to be handled. But he does give a relevant judgment concerning a case of foreknowledge that is a version of the Newcomb Problem [16] (pp. 126–127):

> If we put a human predictor in place of God, and we ask again what would have been the case if I had declined the $1000, the answer will depend on the predictor's modus operandi. First case: the predictor is a time traveler. He saw me accept the $1000, then departed to the past taking his knowledge with him. His foreknowledge is causally downstream from its object. Then I want to hold

> fixed that the time traveler has foreknowledge, and say that if I had declined, the time traveler would have known that I was going to decline . . . .
>
> Second case: the predictor is an expert psychologist, who knows past conditions and regularities of cause and effect. His foreknowledge and its object are separate effects of common causes. Then I want to hold the past fixed, and say that if I had declined, I would have violated some one of the regularities the psychologist relied on.

For Lewis, the crucial difference between the time traveler and the expert psychologist is that the former employs reverse causation to make his pronouncement. Although you should two-box in a Newcomb Problem even with a 100% accurate predictor, in which case you receive $1000, if time travel enables the foreknowledge then you should one-box and receive $1 million. I need to show how this works.

7. A Forking Miracle

Suppose that in the actual world, @ Lewis one-boxes in the time travel Newcomb game. The counterfactual judgment in this case involves a serious back-tracking argument, but the time traveler's "prediction"—unlike the psychologist's—is caused by Lewis's choice [12]. Call the time of return from the future *t2*, and the time of Lewis's one-boxing choice *t3*; then the prediction occurs between *t2* and *t3* (Figure 1).

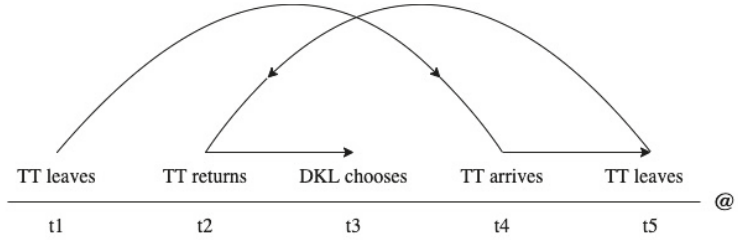


Figure 1. The time-traveling Newcomb game predictor.

The one-boxing Lewis will defend his choice by counterfactual reasoning: *If I had two-boxed at t3, then the predictor would have to have predicted that I would two-box at t3, and there would have to have been no $1 million placed in Box B. So if I had two-boxed at t3, I would have gotten only $1000.*

In the nearest two-boxing world—call it *TB*—it cannot be true that there is exact match of particular fact with @ until just before *t3*, since TB is already different from @ at *t2*, when the time traveler returns from the future with different beliefs. TB need not be different from @ at *t1*, so I shall assume that TB exactly matches @ until just before *t2*, when there is a small miracle.

Now, to the point. In TB, the time traveler's prediction does not cause Lewis's choice of two boxes. Is a second small miracle required, just before *t3*, to produce Lewis's different decision? Suppose that is true. Then by Lewis's metric, TB is a third-rate world, roughly as close to @ as a banana peel world is. By analogy then, Bad-but-Smart Suzy World contains two small miracles: the first just before Bad Suzy's arrival from the future at *t** (say, one day before *t*), and the second just before Bad Suzy's attempt on Baby Suzy's life. But that makes Bad Suzy World only *almost* as close to Good Suzy World as Vihvelin's banana peel world is. Since Lewis at step 2 tells us to *maximize* the region of exact match of particular fact, the banana peel world still apparently wins (with a drastically smaller margin of victory than Vihvelin claims).

But Lewis can do better. Notice that the argument just given for a second small miracle in TB rests on the fact that the prediction does not cause Lewis's choice; but it ignores the existence of reverse causation—the fact that Lewis's choice causes the earlier prediction.

For a centered deterministic world without reverse causation, a standard counterfactual judgment postulates one small miracle, and all the salient divergence from the centered world is causally traceable to the time of that miracle (any other divergence would be gratuitous). But it is misleading to think that the miracle *causes* the divergence, for the miracle is a difference in the laws, not a difference in the pattern of events. Better to say that the small miracle *permits* the divergence in the pattern of events. Then we are free to ask what the pattern of causal relations is between the events in the permitted divergence.

Thanks to the reverse causation, in the nearest world to TB where Lewis two-boxes, one small miracle permits a double divergence in causation. The first salient difference is the difference in the predictor's beliefs at *t2*, and this causes a different prediction, which in turn causes a different prize to be awarded. That and its consequences are one part of the divergence, all of which lie in the external future. In the second part of the divergence, Lewis chooses differently and that has its causal consequences, most of which lie in its external future, but some of which lie in its external past. If it helps, think of the divergence in causation as being doubly present thanks to the reverse causation, by analogy with the way a time traveler can be doubly present by traveling into their own past. When one small miracle permits a double causal divergence, call it a *forking miracle*.

Now, I can give my final judgment on what is true in *Good Suzy*. Had Suzy tried to kill Baby Suzy, she would or might have succeeded, and if she had succeeded someone would have had to have time traveled to past time *t** to arrange the scanning of Baby Suzy, ensuring Baby Suzy's reinstatement after *t*. Thanks to a forking miracle just before *t**, the laws permit the appearance of a time traveler from the future, caused by the (later) killing of Baby Suzy at *t*.

8. It Is Not the Size of the Miracle, It Is What You Do with It

But wait—is the appearance at *t** of a time traveler from out of nowhere not kind of a biggie? Not really, and I do not believe Vihvelin would think so. Lewis's account of his metric should not be read too narrowly. What Lewis calls a *small* miracle Lewis also calls "localized", but it could as well be called a one-off miracle. When making counterfactual judgments we move as far as we need to from the centered world to make the antecedent true, but no further. Any further change would be gratuitous.

In describing the Nixon case, Lewis describes the small miracle that facilitates Nixon's counterfactual pressing of the button as follows [15] (p. 468):

> The deterministic laws of [the centered world] are violated at w_1 in some simple, localized, inconspicuous way. A tiny miracle takes place. Perhaps a few extra neurons fire in some corner of Nixon's brain.

This is potentially misleading, since small miracles don't have to be as small as that one; rather they must be as small as possible to avoid gratuitousness. And if time travel to the past is logically possible, small miracles must permit time travelers to appear in the past in a first-rate counterfactual world at a point earlier than they appeared in the centered world. Suppose that in TB the predictor reliably comes back 15 min earlier with a two-box prediction than it does with a one-box prediction. Then had Lewis two-boxed, the predictor would have had to have arrived from the future 15 min before *t2*. The miracle required to permit that difference does not seem "simple" or "tiny" or especially "inconspicuous".

Lewis's distinction between big and small miracles tempts us to think that a banana peel world is quite like a first-rate world, and quite unlike a fourth-rate world. I see it differently. The metric rules out gratuitous law changes, and in that respect a banana peel world is quite like a fourth-rate world, and quite unlike a first-rate world. The big/small distinction is misleading, because it is not absolute size that matters. Vihvelin's own judgments reflect this. Concerning a centered world such as Good Suzy World, Vihvelin [4] invites us to compare:

(e) If Suzy tried to kill Baby Suzy, she failed.

(f) If Suzy had tried to kill Baby Suzy, she would have failed.

. . . It's not just that (e)—the indicative conditional—is true. (f) also seems to be true.

Stipulate that Vihvelin is correct about (f)—say, because the success would require corpse-resurrection. Corpse resurrection is eliminated at the first step, since it occurs in the set of big miracle worlds. As Vihvelin [4] puts it, these are:

[W]orlds where the baby dies but is subsequently resurrected from the dead and grows up to be the adult Suzy; (These are worlds where, in addition to the small divergence miracle that enables Suzy's attempt, there is a big miracle).

But now consider another pair of conditionals, also evaluated at Good Suzy World:

(e*) If Suzy killed Baby Suzy, Baby Suzy was corpse-resurrected.

(f*) If Suzy had killed Baby Suzy, Baby Suzy would have been corpse-resurrected.

By Vihvelin's account, indicative conditional (e*) is true if corpse-resurrection is required for success. Counterfactual conditional (f*) is true, too, since corpse-resurrection is logically possible. But then, corpse-resurrection must not be a big miracle, else that world would be eliminated at the first step in Lewis's metric. So when evaluating counterfactuals from Good Suzy World, the very same corpse-resurrection world contains a big miracle with respect to (f) but not with respect to (f*). By the same token—as in my treatment of Suzy's counterfactual attempt at autoinfanticide requiring a divergence beginning with a time traveler appearing in the past—even if such appearances would be big miracles in other contexts, that does not show it is a big miracle with respect to *Good Suzy*.

The point about miracle size arises with respect to banana peel worlds as well. Vihvelin postulates a preventative second small miracle, but she does not tell us what it involves. Does the signal from Bad Suzy's brain disappear en route to her arm and hand muscles? That seems, on balance, not a case of an attempt to kill. Given an attempt really carried out—for instance, suppose that Bad Suzy closes her strong grip on Baby Suzy's windpipe, an action that would ordinarily crush it beyond repair—what ordinary occurrence stops her from succeeding? The miracle in question might then need to be quite sizeable, but once again, absolute size is strictly irrelevant to the Lewisian view. What makes a miracle a small miracle is a matter of its not being gratuitous, and here Vihvelin loses the argument.

9. Strange Shackles Indeed

Here is another way to see the same point. Although Vihvelin does not tell us what in particular foils Bad Suzy's attempt, she does say the attempt fails because of the laws of nature [3] (p. 324):

> My arguments support the claim that any time-travel world where any person succeeds in killing the person that is her younger self is a world which includes events that are miraculous by the standards of our laws . . . . So I think we should conclude that the killing of one's younger self is *nomologically impossible* and that is why no one has the narrow ability to do such a thing.
>
> [emphasis original]

Vihvelin then rebuts an objection from Ted Sider [17] that her view after all requires a sort of temporal censor, and that such strange metaphysical "shackles" are better avoided [3] (pp. 324–325):

> I agree with Sider [on the desideratum]. But I deny that my argument commits me to any strange shackles or "exotic metaphysical add-ons".
>
> A time traveler trying to kill her infant self is like a person trying to build a perpetual motion machine . . . . If you try to build a perpetual motion machine you will fail . . . .
>
> And it's not just that anyone's actual attempts have happened to fail, every *counterfactual* attempt *would have failed* as well. If anyone, be they Edison or Elon Musk, had tried to build a perpetual motion machine, they *would have failed*. Not because of exotic metaphysical "forces" or "guardians" or "shackles" but because the creation of such a machine would contradict the laws of thermodynamics . . . .

> Just as the laws of nature entail the impossibility of perpetual motion machines, they also entail that people killed in infancy do not go on to become murderous adults. To suppose that the time traveler could succeed in killing her baby self is to suppose that these laws are false.
>
> [emphases original]

We are now in a position to use Lewis's metric of overall similarity of worlds to give a more nuanced version of Sider's complaint, and show that Sider is correct after all. Vihvelin's argument does postulate metaphysical shackles, and they are strange shackles indeed. Suppose Musk never tries to build a perpetual motion machine. It is true that had he tried, then he would have failed. Assuming determinism, the nearest relevant world is first-rate, and has different laws to ours, but the difference is the one small miracle required to permit the difference in the pattern of events that includes the attempt and its different causal consequences. We should think of a first-rate world as one in which the laws are the same as in the centered world from just after the small miracle. In the general schema, the small miracle occurs just before o. So a possible world which is all and only a duplicate of just the part of w_1 from o onwards has the same deterministic laws as a possible world which is all and only a duplicate of just the part of the centered world from o onwards. Same deterministic laws, different starting conditions, and so a different pattern of events.

The future similarity objection to Lewis's analysis of counterfactuals claims that Lewis's metric of overall similarity must favor a third-rate world which includes a second small miracle: a banana peel that foils Nixon's holocaust attempt, for instance. Notice what we should not even try to say. We should not say that *our* laws require any second small miracle. Quite the reverse: *our* laws *rule out* such a miracle—that is the point of Lewis's reply to the objection. The closest world is a first-rate world because that is a world with the same laws as the centered world, excepting only the small miracle required to permit the antecedent to be true. So Vihvelin is right about Musk's counterfactual attempt at perpetual motion. In a first-rate world he tries and fails, since *our* laws require the failure. He does not fail because of a sui generis banana peel. Musk fails because our laws have the fail-safes built in; they would foil anyone.

But this is manifestly not true of Vihvelin's postulated banana peel world. As we saw in Section 5, fail-safe worlds which would foil almost anyone are not relevant to Vihvelin's argument, and I therefore stipulated that Good Suzy World has no such fail-safes. So it is misleading at best to say as Vihvelin does that our laws—or more precisely, the laws of Good Suzy World—require a second small miracle that foils Bad Suzy's attempt. Instead, Vihvelin must postulate that Bad Suzy World has different laws, laws that foil in Bad Suzy World in a way that the laws of Good Suzy World do not. The laws of Good Suzy World require that there be no such extra miracle. So the extra miracle is indeed an exotic metaphysical add-on, a strange shackle of the sort Vihvelin agrees we should avoid if possible.

10. Conclusions

My response to Vihvelin has been long and involved. That is the nature of the beast. In the final analysis, though, Lewis's position is not only defensible but sensible. Time travel fiction is full of folks who really ought to know better trying to do things that just will not happen. Good Suzy is different, and better, and never even tries to kill her Baby self. Bill and Ted are also different, not trying to change things, but wisely using time travel to bring about desirable results. Bad Suzy is different again, because she by her own unwise actions is forced to emulate Bill and Ted. Like Good Suzy, Bad Suzy should simply have left well enough alone. Like Good Suzy, Bad Suzy should have reasoned: *I **could** kill Baby Suzy, but why would I? Why would I even try?*

Ultimately, no big miracle is needed for Suzy to succeed in killing Baby Suzy; autoinfanticide is no biggie, at least for Lewis. But there is more work to be done. To give a more general defense, I must defend the Lewisian account of survival against its many detractors, but that is a task for another time. Second, I have defended Lewis only given determinism.

I myself am a determinist, but Lewis is not, so a complete defense of Lewis would have to accommodate his application in [18] of the metric to indeterminism. Finally, both Vihvelin and Wasserman [5] independently object to Lewis's metric of overall similarity of worlds even for deterministic laws. I believe Lewis's metric needs some revision, but predict that the revisions will not undermine Lewis's position on what time travelers can do. Again, that is a topic for another time.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 It's unclear whether Lewis here is using "ability" stipulatively or trying to capture the platitudes surrounding its use. Against the latter, it seems we often say we were "unable" to do something we could not do in the "luck" sense.

2 Lewis gives a second reason for the logical impossibility of changing the past in a single timeline, an argument which appeals to atomism about temporal duration, using "moment" as a technical term for a temporal atom [1] (p150). Though I am a fellow atomist, I prefer the more general appeal to flat contradiction.

3 Vihvelin shows that her view generalizes to some other cases involving reverse causation, where non time travelers are similarly restricted [3] (pp. 322-323). I shall not examine those extra cases here.

4 Some defenses even extend to replacing timelines, such as the branching time versions of the Suzy case suggested by John Carroll [6].

5 If Yeva's were the dead-on-arrival case there would only be one continuer that is a person stage; hence it would not be a fission case.

6 It's not simply the amount of time elapsed, however. It's that given normal changes, more time leads to weaker bonds.

7 One often hears that the scanning and replication technology is nomologically impossible because it violates Heisenberg Uncertainty. At most, this shows that it's impossible to *exactly* replicate. But since ordinary survival does not require that successive stages be intrinsic duplicates, neither does relocation or reinstatement. Whatever is near enough for ordinary survival will be good enough for relocation or reinstatement.

8 To be fair, Lewis at one point describes a relocation scenario as "far-fetched" [10] (p. 192, n. 3), but that is a very different case in which the scanning is somehow done remotely, à la *Star Trek* "beaming".

9 At least, according to Lewis's *Analysis 1*, which assumes that the background truths in fiction are supplied by actual world facts. I shall for simplicity ignore his alternative analysis which appeals instead to a background of mutual shared belief. (Analysis 1 is far more plausible, in any case).

10 Note well that neither Lewis nor Vihvelin is postulating worlds where the laws of *that world* are broken. A miracle big or small is instead a difference between the laws of different worlds.

11 There are other ways the story could go, all dependent upon the fact that Bad Suzy World is a time travel world. Perhaps one of Bad Suzy's parents would not have maintained their anti-reinstatement stance when confronted with Baby Suzy's corpse, especially given Bad Suzy's presence; and hence might be their own daughter's kidnapper. What is certain is that *somebody* does what is needed.

12 Given determinism *every* time-indexed standard counterfactual resolution involves a back-tracking argument, since there had to be a small miracle before the time in question; call these back-tracking arguments *minor*. A serious back-tracking argument takes us back before a different, earlier time index. The presence of a serious back-tracking argument is not sufficient for a backtracking counterfactual resolution; the resolution delivered must also differ in truth value from the "standard" resolution [15] (p. 457). I think Lewis should say that since there is reverse causation the one-boxer's serious back-tracking argument *delivers* the standard resolution.

References

1. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
2. Vihvelin, K. What Time Travelers Cannot Do. *Philos. Stud. Int. J. Philos. Anal. Tradit.* **1996**, *81*, 315–330. [CrossRef]
3. Vihvelin, K. Killing Time Again. *Monist* **2020**, *103*, 312–327. [CrossRef]
4. Vihvelin, K. Counterfactuals, Indicatives, and What Time Travelers Can't Do. Available online: www.vihvelin.com (accessed on 8 September 2021).
5. Wasserman, R. *Paradoxes of Time Travel*, Illustrated ed.; Oxford University Press: Oxford, UK, 2018.
6. Carroll, J.W. Ways to Commit Autoinfanticide. *J. Am. Philos. Assoc.* **2016**, *2*, 180–191. [CrossRef]
7. Vranas, P.B.M. What Time Travelers May Be Able to Do. *Philos. Stud. Int. J. Philos. Anal. Tradit.* **2010**, *150*, 115–121. [CrossRef]
8. Lewis, D. *On the Plurality of Worlds*; Blackwell: Oxford, UK, 1986.

9. Lewis, D. Survival and Identity. In *The Identities of Persons*, Revised ed.; Rorty, A.O., Ed.; University of California Press: Berkeley, CA, USA, 1976; pp. 17–40.
10. Lewis, D. Academic Appointments: Why Ignore the Advantage of Being Right? Chapter. In *Papers in Ethics and Social Philosophy*; Cambridge Studies in Philosophy; Cambridge University Press: Cambridge, UK, 1999; Volume 3, pp. 187–200. [CrossRef]
11. Lewis, D. How Many Lives Has Schrödinger's Cat? *Australas. J. Philos.* **2004**, *82*, 3–22. [CrossRef]
12. Lewis, D. Postscript A to "Survival and Identity". In *Philosophical Papers Volume I*; Oxford University Press: Cambridge, UK, 1983; pp. 73–76. [CrossRef]
13. Lewis, D. Causation. *J. Philos.* **1973**, *70*, 556–567. [CrossRef]
14. Lewis, D. Truth in Fiction. *Am. Philos. Q.* **1978**, *15*, 37–46.
15. Lewis, D. Counterfactual Dependence and Time's Arrow. *Noûs* **1979**, *13*, 455–476. [CrossRef]
16. Lewis, D. Evil for Freedom's Sake? *Philos. Pap.* **1993**, *22*, 149–172. [CrossRef]
17. Sider, T. Time Travel, Coincidences and Counterfactuals. *Philos. Stud. Int. J. Philos. Anal. Tradit.* **2002**, *110*, 115–138.
18. Lewis, D. Postscript D to "Counterfactual Dependence and Time's Arrow". In *Philosophical Papers Volume II*; Oxford University Press: Cambridge, UK, 1987; pp. 58–65. [CrossRef]

philosophies

MDPI

Article

Does Lewis' Theory of Causation Permit Time Travel?

Phil Dowe

School of Philosophy, Australian National University, Canberra, ACT 0200, Australia; phil.dowe@anu.edu.au

Abstract: David Lewis aimed to give an account of causation, and in particular, a semantics for the counterfactuals to which his account appeals, that is compatible with backwards causation and time travel. I will argue that he failed, but not for the reasons that have been offered to date, specifically by Collins, Hall and Paul and by Wasserman. This is significant not the least because Lewis' theory of causation was the most influential theory over the last quarter of the 20th century; and moreover, Lewis' spirited defence of time travel in the 1970s has shaped philosophers' approach to time travel to this day.

Keywords: time travel; counterfactuals; causation; miracles

Citation: Dowe, P. Does Lewis' Theory of Causation Permit Time Travel? *Philosophies* **2021**, *6*, 94. https://doi.org/10.3390/philosophies6040094

Academic Editor: Alasdair Richmond

Received: 14 September 2021
Accepted: 17 November 2021
Published: 23 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction: Lewis' Theory, as Advertised

In 'Causation' [1], David Lewis presents the theory that *c* causes *e* if *c* and *e* occur and 'had *c* not occurred, e would not have occurred' is true (='counterfactual dependence'), or if there is a chain of such counterfactual dependence connecting *c* and *e*. The counterfactual 'had *c* not occurred, *e* would not have occurred' is true iff *e* does not occur in any of the closest ~*c* worlds. This analysis is intended to work under determinism. In that 1973 paper, Lewis says " [We should not reject] a priori certain legitimate physical hypotheses that posit backward or simultaneous causation" [1] (p. 566). I will argue that despite Lewis' best intentions, he has unwittingly done so. I will also show that the argument applies equally to Lewis' final version [2] of the counterfactual theory.

In 'The Paradoxes of Time Travel' [3], Lewis defends the possibility of time travel on the basis of an eternalist metaphysics of time and a purdurantist–causal theory of persistence. Since two stages of a time traveler need to be connected by appropriate causal connections in order for the two stages to be part of the one person, and hence to be a time *traveler*, "... travel into the past necessarily involves reversed causation" [3] (p. 147). In this paper, Lewis says, 'Elsewhere I have given an analysis of causation in terms of chains of counterfactual dependence, and I took care that my analysis would not rule out casual reversal a priori" [3] (p. 148). I will argue that despite significant effort on Lewis' part, it turns out that he did not take enough care.

In 'Counterfactual Dependence and Time's Arrow' [4], Lewis spells out a more detailed account of comparative overall similarity. His celebrated 'similarity measure' orders worlds as follows:

(A) It is of the first importance to avoid big, widespread, diverse violations of law.
(B) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
(C) It is of the third importance to avoid even small, localized, simple violations of law.
(D) It is of little or no importance to secure approximate similarity of particular fact, even in matters which concern us greatly [4] (p. 472).

Lewis claims that a major advantage of this analysis over alternatives (e.g., [5]) that hold the past fixed by 'brute force' is that his account allows for backwards causation. Lewis adds "Careful readers have thought they could make sense of stories of time travel... speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves. Most or all of these phenomena

would involve special exceptions to the normal asymmetry of counterfactual dependence. It will not do to declare them impossible a priori" [4] (p. 464). I will argue that Lewis, in effect, has done exactly that.

For ordinary forwards causation, according to Lewis' similarity relation, the usual structure of the relevant $\sim c$ worlds where event c is an actual cause of event e will be: perfect match with actuality up until a time t_{c-} just before c; a small miracle leading to $\sim c$; no further miracles; and future divergence including $\sim e$. Specifically, those worlds will be closer than worlds like this: perfect match after t_{c+} including e, numerous diverse miracles leading back to $\sim c$, no other miracles, and past divergence. Thus, to have backwards causation of an earlier event e by c, the idea is that we should look at worlds like this: perfect match after t_{c+}, a small miracle leading back to $\sim c$, no other miracles, and past divergence including $\sim e$. Lewis says:

> "I think I can argue (but not here) that under my analysis the direction of counterfactual dependence and causation is governed by the direction of other de facto asymmetries of time. If so, then reversed causation and time travel are not excluded altogether, but can occur only where there are local exceptions to these asymmetries" [3] (p. 148).

Lewis claims that on his analysis, the direction of counterfactual dependence and causation is governed by what he calls the 'asymmetry of overdetermination'. A 'determinant' is a minimal set of conditions jointly sufficient, given the laws of nature, for the event in question. The designation 'asymmetry of overdetermination' is ambiguous. On the one hand, it refers to a global asymmetry, and on the other hand it refers to an extrinsic property of a particular event, the putative cause. I will use the term exclusively in the latter sense. The property concerns how the event connects lawfully to its surroundings. An event is overdetermined if it has more than one determinant. A particular event, say c, displays an asymmetry of overdetermination if the overdetermination in one direction in time is significantly greater than its overdetermination in the other direction in time. For forwards causation, the cause c must exhibit a significantly greater overdetermination in the future than in the past, and vice versa for backwards causation [4] (p. 474).

On Lewis' account, causation supervenes on the actual distribution of particulars, despite the inter alia reference to possible worlds. Thus, a non-standard way to think about Lewis' counterfactual theory of causation is to note some of its distinctive entailments about actuality: (1) A cause is a nomically necessary condition, given the actual circumstances, for its effect, where 'nomically' means according to the laws, which for Lewis are given by the best system analysis of the actual distribution of particulars; and (2) a cause exhibits asymmetry of overdetermination in the direction of the effect, that is, toward the past or toward the future depending on whether the effect is respectively in the past or in the future of the cause. Both conditions are necessary for causation. The reason (2) is necessary for causation is that, if a cause had no such asymmetry, that is it had the same number of determinants in the future as in the past, then a world with a perfect match in the past up to a small miracle leading to $\sim c$, would be no closer than a world with a perfect match in the future back to a small miracle leading to back to $\sim c$. In the former world, there would be no event e, in the latter there would be, and since e must be missing in all the closest $\sim c$ worlds for c to cause e, it follows that an event which lacks asymmetry of overdetermination cannot be the cause of anything. Actually (2) cannot be quite right, as I shall later illustrate, because it is defined in terms of the distinction between the past and the future of the event in question, whereas the similarity relation does no such thing, appealing only to 'spatiotemporal regions'.

As I have indicated, contrary to advertisement, Lewis' account fails to allow for backwards causation. I will sandwich my main argument for this (Section 3) between discussions of some specious arguments for the same conclusion.

2. Specious Argument #1

Suppose c is the cause of an earlier event e, that c has earlier causes including b, and also later effects including d. For e to be an effect of c, c must exhibit an asymmetry of overdetermination towards the past. Suppose the following world is closer than any $\sim c$ world containing e: perfect match after t_{c+}, small miracle leading back to $\sim c$, no other miracles, and past divergence including $\sim e$. However, that world contains d, so it follows that c cannot also be a cause of d; but that is to be expected, because for d to be an effect of c, c must exhibit an asymmetry of overdetermination towards the future. Since it is logically impossible for an event to exhibit an asymmetry of overdetermination in both the past and future directions, no event can have both a past and a future effect. Yet typical time travel scenarios, and the 'speculative physicist's' hypotheses mentioned by Lewis, do involve events with past and future effects. Therefore, Lewis does not allow for typical time travel or backwards causation.

This is a specious argument. The similarity relation does not trade on the distinction between past and future—it is formulated in terms of spatiotemporal regions. Space-time divides, for example, into three regions and the similarity relation rules on the relative closeness of a world where part of the past matches actuality, and part of the past diverges (Compare [6] p. 10). Such a world—with reference to the above case—might be: Perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future including $\sim d$, and divergence in the past in a region stretching from $\sim c$ back through a region including $\sim e$. Then indeed, c causes both e and d. The condition for causation implicit in the similarity relation is not 'a cause exhibits asymmetry of overdetermination in the direction (past or future) of the effect', but rather, (2)* a cause exhibits asymmetry of overdetermination in the spatiotemporal region containing the effect compared to some other region very close to c. To reject this point is to accept the specious argument, I would claim, and among other things that would leave one wondering how Lewis ever thought his account could get off the ground.

This brings us to an exception that should be allowed to the argument I will put below. I will discuss the exception first, then give the argument in the next section. Suppose again that c is the cause of an earlier event e, that c has earlier causes, including b, as well as later effects. Call the region containing all of c's past effects the 'time travel region of c', and call the region containing all of c's past causes the 'causal past of c'. Now we make two assumptions. (1) Suppose the time travel region of c and the causal past of c do not overlap, or more precisely, there exists regions R, S, s.t. R contains all of c's past effects, S contains all of c's past causes, and R and S do not overlap. (2) Suppose, more generally, nothing in the time travel region of c causes anything in the causal past of c. For any case where these assumptions hold, the closest world will indeed be: perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future, and divergence in the time travel region of c including $\sim e$. Thus, c causes e, and Lewis' analysis does indeed permit time travel. The assumptions are very restrictive, perhaps too restrictive to be of interest, but it does give us an exception to the argument I am now going to give.

3. The Main Argument

Suppose I want to see a comet pass near the earth, but I sleep through the alarm[1]. Suppose someone (my older self) subsequently gives me instructions for a time machine (event b). I find the time machine, set the controls as instructed, and press the button (c). The machine immediately travels back in time; nevertheless, c has future effects, including the sound of the button being pushed reaching (d) the ears of a nearby possum. I then get to see the comet (e), and then I find my younger self and give him the instructions (b). c causes e, a case of backwards causation (Figure 1). Suppose the scenario just described plays out in the actual world, $W_@$. Let us make two assumptions, both purely heuristic: (1) nothing earlier than e has a future cause, and (2) no future event (including c) has any effects in the region of b except via the region of e. It does not matter for the following argument

whether the time travel occurs via closed timelike curves, which does not involve any local backwards causation, or via a 'Wellesian' time machine, which does.

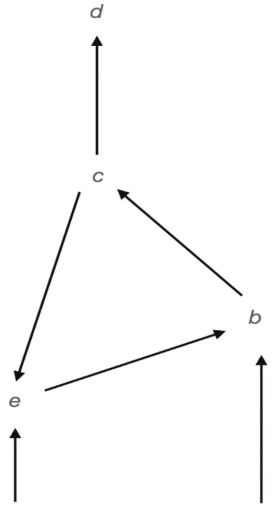


Figure 1. If I had not pressed that button?

I will start by asserting which worlds are among the closest worlds[2], then I will show that if those are the closest worlds, then c does not cause e, and then I will show that those worlds are indeed (in general) the closest worlds. The closest $\sim c$ worlds are worlds like W_1 and W_2, as follows:

W_1: Perfect match up to t_{e-}, thereafter divergence, no miracles, $\sim e$.
W_2: Perfect match up to t_{e-}, thereafter divergence, no miracles, e.

(I say, 'Perfect match up to t_{e-}', but this perfect match may extend a little further into the future of e in the region of the immediate past of b, depending on what one means by 'future'.) An example of W_1 would be: I sleep through the alarm, no-one appears out of a time machine, no-one gives me the instructions, I do not even know about the time machine, I do not time travel, and I do not get to see the comet. I simply regret sleeping through. This is a $\sim c$, $\sim e$ world. An example of W_2 would be: I sleep through the comet, then someone (her older self) subsequently gives my wife instructions for a time machine. She takes me to the time machine, sets the controls as instructed, and she presses the button. We both time travel, I see the comet, and she passes the instructions to her younger self. This is a $\sim c$, e world; that is, it is not true that I press the button, but it is true that I see the comet. In neither of these worlds are there any miracles. The past of e and b matches actuality, but what ensues depends on not only on that past, but also on what happens in the future, including whether or not the button is pressed. Vary the latter, and you vary what happens in the region of e and b, without miracles. These are the closest worlds, and if I am right about that then c does not count as a cause of e on Lewis' theory, since his theory requires that e does not occur in any of the closest $\sim c$ worlds. W_2: is one of the closest $\sim c$ worlds, e occurs in W_2, hence c does not cause e.

In fact, in general, there would be indefinitely many such closest worlds, each containing no miracles. This feature is well-known in the literature on wormhole time machines in the context of general relativity [7–9], but is not so well-appreciated in the philosophical literature concerning Wellesian time travel. In the physics literature, it has been shown for certain classes of cases that data fixed before a time travel region can be extended to indefinitely many consistent trajectories through the time travel region [7]. Thus, the general theory of relativity is, in one sense, an indeterministic theory (although not in another sense,

since all the local dynamics will be deterministic). What happens in the region of *e* depends deterministically on what happens before *e together with* what happens in the region of *c*. On the other hand, there is indeterminism in the sense that what goes on before the time travel region does not fix what occurs in the time travel region [8]. Laplacean determinism (i.e., that the particulars at a given time together with the laws fixes the particulars at later times) fails. This, however, does not enable us to define chances in the usual way, so it is appropriate that we seek to use Lewis' theory designed for determinism. In any case, Lewis' theory designed for indeterminism, where causation obtains on account of counterfactual chances, was not developed far enough to see how Lewis might account for backwards causation (see Lewis' comments at [10] (p. 274), for a discussion of the constraints on chances in loops see [11]).

Assuming still that I am right about the closest worlds, why does Lewis' theory go wrong? It is not because there is anything wrong with the idea that causes are necessary conditions for their effects in the actual circumstances. The problem in time travel cases is that Lewis' similarity relation fails to do something specific that it is intended to do: it fails to pick out the worlds in which the actual relevant background conditions are held fixed. If you allow the background conditions to vary, then in general, a cause will not be a necessary condition for its effect, even if it is a necessary condition for its effects in the actual circumstances. The reason it fails to pick out the worlds in which the actual relevant background conditions are held fixed is that large miracles are required to hold fixed the background of the cause on account of the fact that the cause has effects in that region. The asymmetry of overdetermination makes it impossible to avoid that, short of conditions that separate the regions containing the causal past and regions containing the backwards effects. That is why some of our closest $\sim c$ worlds contain *e*.

So let us turn to the question of whether worlds like W_1 and W_2 are indeed (in general) the closest worlds according to Lewis' similarity relation. Worlds like W_1 and W_2 contain no miracles. A world with more perfect match gained at the cost of a large miracle will not be as close as W_1 and W_2. The only way to get a closer $\sim c$ world is to buy increased perfect match at the cost of a small miracle or two. Let us try that. What we would like to do in our example is to have perfect match in the region of *b*, and a small miracle leading to $\sim c$. Indeed, such a world can have divergence in the future and in the region of *e*, such that we get the required result that *e* and *d* do not occur at that world. The problem is, in such a world we need a large miracle to go from divergence in the region of *e* to perfect match in the region of *b*, if Lewis' similarity relation is right about the asymmetry of overdetermination. Each and every change in the *e* region which would have effects in the *b* region each of which would need to be somehow deleted by a miracle in order to maintain perfect match in the *b* region, and that set of diverse miracles would add up to a large miracle. Alternatively, a closer world (W_3 below) than the one just described is one where the reconvergence miracle occurs between *c* and *e*, not in the region between *e* and *b*, and this would be a world containing *e*. Both of these worlds contain a large miracle, and hence, neither is as close as W_1 or W_2.

> W_3: Perfect match up to t_{e^-}, perfect match in the region of *b*, small miracle leading to ~c, future divergence, large reconvergence miracle, *e*.

It may be objected that I have only shown that there is no counterfactual dependence of *e* on *c*, not that there is no causation given that on Lewis' theory causation is the ancestral of counterfactual dependence. However, it is easy to show that there can be no chain of counterfactual dependence either. The first link of any chain of counterfactual dependence between *c* and *e* must be a counterfactual dependence of something, call it *f*, on *c*. For *f* to counterfactually depend on *c* it must be that none of the closest $\sim c$ worlds contain *f*. These $\sim c$ worlds are exactly the ones we have already been considering. The closest are worlds like W_1 and W_2. By the reasoning above, some of these worlds contain *f* and some do not. Hence, there is no counterfactual dependence of *f* on *c*, and hence, there can be no chain of counterfactual dependence between *c* and *e*, hence *c* does not cause *e*.

Alternatively, it may be objected that I have only shown that there is no backwards causation on Lewis' original theory, but not for his final theory [2]. On this theory *influence* is "a pattern of dependence of how, when and whether upon how, when and whether" [2] (p. 190), and causation is the ancestral of influence. This allows for more cases to count as causation compared to the original theory, namely, all those cases which do not exhibit whether on whether dependence yet do exhibit how, when or whether on how, when or whether dependence. In fact, influence can obtain when any one of nine kinds of dependence obtains, whether on whether being one of those. So to show that there is no whether on whether dependence (counterfactual dependence) in a case of putative backwards causation, as I have done, is not to show that there is no influence. However, it is straightforward to show that the argument I have given applies equally to any case of influence. On this theory, causation obtains if at least one of the nine types of dependence obtains. But for any type of dependence to obtain, the defined change ('alteration') must feature in *all* the relevant closest worlds, just as for counterfactual dependence. Lewis' semantics for such counterfactuals remains the same. Thus, my argument will apply to each kind of dependence. Take when on when dependence. Counterfactually, suppose by a small miracle the button is pushed slightly later (c^*), and that the time travel set-up is such that I see the comet slightly later (e^*). However, by the argument given above, closest c^* worlds are worlds with no miracles and no perfect match in the time travel region, and therefore some of those worlds do not contain e^*. The same argument applies to any alteration to the time of the cause. Hence, there is no when on when dependence. A similar case can be made for each of the nine patterns of dependence, hence c does not influence e. Hence, Lewis' influence theory does not allow for backwards causation.

Return now to the two assumptions of the argument, namely, (1) nothing earlier than e has a future cause, and (2) no future event (including c) has any effects in the region of b except via the region of e. Would it make any difference if we relaxed these assumptions? No. Take assumption (1). This would hold for certain events around a time machine wormhole: events near enough the so-called Cauchy horizon [7]. In fact the Cauchy horizon could be characterized roughly as the set of points which divides the region which can contain events with future causes (time travel region) from the region which cannot (Cauchy region). However, in the Wheeler-Feynman advanced/retarded potential theory assumption (1) would not hold as backwards effects would extend indefinitely into the past, with intensity dropping off with $1/r^2$. Suppose in our example e does have past effects (i.e., assumption (1) is false). If these effects extend indefinitely into the past, then the closest $\sim c$ worlds will be worlds with no miracles but no perfect match—these worlds are still closer than any world with a large miracle. Suppose on the other hand that although e does have past effects so assumption (1) is false, nevertheless there is some earlier event a which is an effect of e but which itself has no earlier effects and indeed that assumption (1) holds of a. Then the closest $\sim c$ worlds would be worlds with perfect match up to a time just before a, and no miracles. Thus the argument still holds if we drop assumption (1).

What about assumption (2)? Suppose c for example has direct effects throughout the region of b, that is, effects which are not 'mediated' by events in the region of e. For simplicity retain assumption (1). Then again, the closest $\sim c$ worlds will be worlds like W_1 and W_2. Now, to get perfect match in the region of b we not only need a large miracle to remove the backward effects in the region of e due to changing c, we also need a large miracle to remove the direct backward effects in the region of b due to changing c. No reason here to doubt my argument.

Assumption (2) does not hold in standard examples of time travel. Lewis' example of Tim who tried to kill his grandfather involves a certain spatiotemporal region replete with both the effects and the causes of future events. Tim's aiming his gun at Grandfather is the effect of Tim's later decision to travel back to kill him. Tim's Grandfather surviving is the cause of Tim's later decision to travel back to kill him. But again, in order to hold fixed the background condition of Tim's pressing the time travel button, we need large miracles to erase the effects of his counterfactually not doing so.

Finally, consider again the exception I granted at the start of this section. In this case, we made two assumptions: (1) the time travel region of c (the region containing all of c's past effects) and the causal past of c (the region containing all of c's past causes) do not overlap, and (2) nothing in the time travel region of c causes anything in the causal past of c. I said the closest worlds will be: perfect match in the region of b, up to a small miracle leading to $\sim c$, divergence in the future, and divergence in the time travel region of c including $\sim e$. This entails that there are no causal loops. So, should the constraint on my argument simply be that there are no causal loops? No, suppose there are 'epiphenomena' of c in the regions of e and b, where epiphenomena are events with no effects and the epiphenomena of c are the effects of c in with no effects. Suppose—as in the exception I granted—that there are no causal loops. Then again, to get perfect match in the region of b in a $\sim c$ world, we need a large miracle to remove the epiphenomena of changing c. Thus the exception cannot simply be granted to worlds with no causal loops.

We now recap the argument. Time travel hypotheses and other backwards causation hypotheses generally require regions containing both effects and causes of some pertinent event, say c. Under such circumstances Lewis' semantics rule out c being the cause of anything in that region (except for causes, if there are any, that are necessary for their effect in all physically possible circumstances). Hence, there can be no time travel on Lewis' theory except where there is no such spatiotemporal and causal overlap.

4. Specious Argument #2: Collins, Hall and Paul

Collins, Hall and Paul [12] (pp. 9–11) give a different argument to the conclusion that Lewis' semantics does not allow backwards causation. In their example a single billiard ball approaches a Time Machine Portal, entrance on the right, exit on the left. It is on course to pass harmlessly between the entrance and exit but a ball emerges from the exit at $t = 0$, collides with the first ball and sends it into the entrance at $t = 1$. Surprise, it is one and the same ball. Suppose:

c—the ball rolling into the entrance
e—the ball emerging from the exit

"Clearly, c causes e", they say [12] (p. 11). So which are the closest $\sim c$ worlds? Consider these $\sim c$ worlds:

W_1: Small miracle leading to $\sim c$ (ball vanishes before rolling into the entrance); a small miracle to re-instate the ball leaving the exit; e; perfect match except for the small region where the ball has disappeared.
W_2: Perfect match in the past, the ball goes straight through with no collision; no ball through the time machine; $\sim e$.

W_1 closer than W_2 because it has the greater region of perfect match which trumps the fact it has more (two) small miracles. Therefore c does not cause e.

We must reject this analysis. The problem is that it trades on an example which is 'simple and ideal'. The example contains just one particle which could be, just as well for the example, a Newtonian point particle. But it can be shown that Lewis' semantics does not apply to simple ideal worlds (compare [13]); causation, on Lewis account, only applies in systems with sufficient complexity. This point will be established in Section 5.

To be convinced at least that the example trades on being 'simple and ideal', one just has to add 'sufficient complexity'. Suppose we have a billiard ball, and suppose there is a strong light source so that there is plenty of light being continually reflected off the ball. Suppose in addition there is a rough surface which results in a noise when the ball rolls over it. Suppose for simplicity we have a short-lived wormhole, so that there is a Cauchy horizon at t_{cauch}. Light will of course enter the time machine, and thereby be propagated back near but not beyond t_{cauch}. Suppose again a ball is on course to pass harmlessly between the entrance and exit but a ball emerges from the exit at $t = 0$, collides and sends the ball into the entrance at $t = 1$. Again, it is one and the same ball. This time, however, if we remove c, it would take a large miracle to regain perfect match anywhere in the

time travel region. Say we have again a small miracle to delete the ball as it rolls into the entrance, and another small miracle to reinstate it leaving the exit. The ball is missing through the wormhole, and any light that would have been reflected off the stage now missing will also be absent, as will any sound waves that the missing stage of the ball would have initiated. We could for example begin to recover perfect match everywhere by a large number of miracles which 'reflect' the light as if it were reflecting off the ball (where it is missing). This might occur by numerous small miracles producing photons with the right properties one by one; and in the same way removing the light that actually travels straight past (where the ball would have been); together with miracles to generate the noise the ball would have made. This array of small miracles adds up to a large miracle. So compare:

W_1: A small miracle leading to $\sim c$ (ball disappears), a small miracle to re-instate the ball at exit, e; a large miracle to ensure perfect match.
W_2: Perfect match in the past until t_{cauch}; ball goes straight through, no collision; no ball through the time machine, $\sim e$.

W_2 is closer than W_1 since the latter contains a large miracle.

However there are other $\sim c$ worlds which do contain e, and which are equally close to W_2. For example, there are closest $\sim c$ worlds containing 'lions'—loops objects which have no beginning or end[3], and in particular have no presence before t_{cauch}. In our case the 'lion' could be a second billiard ball, but let us stick with the literal lion. Consider this $\sim c$ world:

W_3: Perfect match in the past until t_{cauch}; no miracles; lion picks up the ball and takes it through the time machine, releases it so it comes out as before, e.

W_2, W_3 are equally close worlds; in fact they are among the closest worlds and both are closer than W_1. Thus c does not cause e. However, now we see the argument conforms to the argument of Section 3 and does not rely on the fact that components of the set up are simple and ideal. Thus, the analysis of Collins et al. is the wrong analysis: Lewis fails to account for backwards causation in their example because their example does not have sufficient complexity for there to be any causation, backwards or otherwise. We will now defend that premise.

5. Lewis' Emergent Causation

"Overdetermination in Lewis' sense . . . "fades out" as we approach the micro level", says Huw Price [13] (p. 150). Suppose the actual world contains just two ideally elastic inert point particles, which collide just once. Suppose c is a certain instantaneous state of the first ball at a point before the collision, and e is a certain instantaneous state of the second ball at a point after the collision. We want to say c causes e. What are the closest $\sim c$ worlds? Compare:

W_1: Perfect match up until t_{c-}; small miracle leading to $\sim c$; no further miracles; future divergence; $\sim e$.
W_2: Perfect match after t_{c+}; small miracle leading back to $\sim c$; no other miracles, past divergence; e.

W_1 and W_2 are to be taken as equally close, hence c does not cause e. This arises because there is no asymmetry of overdetermination at the level of the simple and ideal, at least when you have time-symmetric dynamics. The asymmetry, as pointed out above, is necessary for causation. However, there are worlds closer than either W_1 or W_2. Compare:

W_3: Perfect match up until t_{c-}; small miracle leading to $\sim c$; small miracle leading to the production of an identical particle *ex nihilo* at t_{c+}; no further miracles; perfect match after t_{c+}; e.

W_3 is closer than W_1. It is worth a second small miracle to purchase a large swath of perfect match, according to Lewis' similarity measure. Hence, c does not cause e.

So there are two reasons there can be no causation on Lewis' theory at the level of the simple and ideal. The first is that following the counterfactual absence of an actual

event a world can reconverge to actuality without a large miracle. The second is that an event can appear without causal precedence and without a large miracle, thereby buying additional perfect match. Once we have sufficient complexity and (presumably, although the connection is quite controversial) an entropy gradient like ours, then it will take a large miracle to reconverge in the first case, and in the second case although it still just takes a small miracle for miraculous replication, now that will not buy any additional perfect match, and so will be ruled out on account of having an extra miracle, on Lewis' similarity measure. In addition, when we consider large scale events or objects, there must be some point at which a miraculous replication requires a large miracle. The miraculous appearance of anything with a considerable number of diverse parts will surely count as a large miracle. For Lewis' theory of causation, then, causation is emergent: it arises only where there is sufficient complexity for events to exhibit the asymmetry of overdetermination.

6. Specious Argument #3: Wasserman

Wasserman [16] gives an argument somewhat similar to that of Collins, Hall and Paul in that it also trades on the simple and ideal, but different in that it is based on the phenomenon of action at a temporal distance. In his example a single electron persists from t_1 to t_3, at which point it enters a time machine. At that moment, a button is pressed and the electron is sent back, discontinuously, to t_2, where it continues on to t_4 and on into the future. Assume that no other electron appears at t_2, ie there are no preempted backups. [16] (pp. 143–144). Wasserman claims this counterfactual is true: 'If the button hadn't been pressed at t_3, an electron wouldn't have appeared at t_2'.

A world W_1 where a small miracle leads to the appearance of an electron at t_2—call this miraculous replication—and a second small miracle leads to the button not being pressed is closer than a world W_2 with no appearance of an electron at t_2 and a single small miracle leading to the button not being pressed. The latter has perfect match only up until t_2 while the former has perfect match up until t_3, *viz.*:

> W_1: perfect match until t_3; a small miracle; appearance of an electron at t_2; second small miracle at t_3; the button is not pressed.
> W_2: perfect match until t_2; no electron at t_2; a small miracle at t_3; the button not pressed.

Therefore it is false that if someone hadn't pressed the button at t_3, an electron wouldn't have appeared at t_2. And so, on Lewis' account, pressing the button is not the cause of the electron landing back at t_2. Further, the culprit can be identified, Wasserman claims:

> Note that this line of reasoning does not apply in the case of ordinary future-directed counterfactuals . . . In the case of time travel, the traces of the button-pressing are irrelevant, since those traces do not disrupt perfect match in the past. This is the fundamental problem for Lewis—in the case [of] backward counterfactuals, perfect match can be purchased without a miraculous cover up. That is why we get the wrong results . . . [16] (p. 147).

A straightforward counterargument shows that it is not only in cases of backwards counterfactuals that we meet this problem. Assume a Newtonian spacetime which allows action at a spatial distance. Suppose when the button is pressed at p_3 (now using p for a spacetime point) an electron is teleported instantaneously to some far away location p_i. As with Wasserman's example there are closer no-button-pressing worlds than the world with perfect match in the past until by a small miracle the electron disappears just before p_3, thence no perfect match or miracles, and no appearance of an electron at p_i. One such closer world would be a world with perfect match in the past until by a small miracle the electron disappears just before p_3, but then by a small miracle an electron appears at p_i. As with Wasserman's example, this world exhibits a greater region of perfect match at the cost of a second small miracle, assuming only that any other effects of the button pressing do not reach that region until some time later. However, in response Wasserman can easily adjust the diagnosis to encompass simultaneous causation (or for that matter causation

outside the lightcone in a Minkowski spacetime). Again, perfect match can be purchased without a miraculous cover up.

The reasoning can be extended to a limited set of forwards in time cases. Suppose in our Newtonian spacetime the forward lightcone limits causation except for rare action at a distance. Then if the button pressing sends the electron to some future location outside the forward light cone of the button pressing, we again find miraculous replication in the closest worlds rules against causation. Lewis' account fails to account for (a limited kind of) forwards in time causation.

The correct diagnosis is that Wasserman's example trades in the simple and ideal, just like that of Collins et al.. As I have argued, on Lewis's account causation is emergent at higher levels of complexity. The reason "perfect match can be purchased without a miraculous cover up" in the appearance of an electron ex nihilo derives from the 'simple and ideal' nature of the micro events and so that we shouldn't expect there to be causation on Lewis' theory, quite apart from any backwards causation. According to Wasserman there is an implausible asymmetry entailed by Lewis' account [16] (p. 147). If we supposed a person rather than an electron was sent back in time then we wouldn't have the problem he gives, because it would take a large miracle to bring forth a person *ex nihilo* at t_2. Such a world would not be a closest world, and hence pressing the button comes out as the cause of the person landing back at t_2. I agree it would take a large miracle to bring forth a macro object *ex nihilo*. I also agree there is an asymmetry between the two cases, but not because one is a case of causation and the other is not; but because there are different reasons in each case for why there is no causation.

It is not the case on Lewis' theory that pressing the button comes out as the cause of a person landing back at t_2. The reason is not that a small miracle could be responsible for miraculous replication of a person, but that it could happen without any miracle. Suppose actually I pressed the button and showed up discontinuously at an earlier time t_2. Consider the worlds where I do not press the button. One such world is where I do not show up at t_2. Another is where I do not press the button, I go home, someone appears from the future, convinces me to go back to the machine, presses the button for me, and then I show up discontinuously at the earlier time t_2. These two worlds are equally close: perfect match until t_2; one small miracle. One has the putative effect, the other does not; so on Lewis' theory, my pressing the button is not the cause of my turning up at the earlier time t_2. This is essentially the same problem as the one I outlined in Section 3. Anything can happen in the time travel region.

But there is another feature of Wasserman's example in addition to the time travel and the simple and ideal aspect, that might be considered to raise a red flag: spooky action at a distance. Should we identify the problem in this example with the spooky action at a spatiotemporal distance, rather than with the time travel or the simple and ideal aspect? Action at a temporal distance is problematic in its own right [17], although Lewis wants to allow for the possibility [3] (p. 148). If we disallow action at a distance then there is no putative counterexample here to backwards causation. Wasserman thinks otherwise: "I focus on discontinuous time travel for the sake of simplicity. The same point can be made in the case of continuous time travel" [2] (n. 11, p. 149). This appears to be an error. Suppose on pressing the button at t_3 the electron turns around in time and travels back continuously to t_2 where it turns around and travels forwards in time[4]. We can consider the world where a small miracle leads to the appearance of an electron at t_2 and a second small miracle leads to the button not being pressed, but this does not generate any further perfect match, since there is no electron traveling backwards between t_3 and t_2 whereas in the actual world there is. On account of its second miracle, that world is not as close as any world with no appearance of an electron at t_2 and a single small miracle leading to the button not being pressed. However, again there are other equally close worlds where an electron does appear without a miracle—some other use of the time machine could be responsible provided only that it does not count as the event pressing the button at t_3. Hence the continuous example does not count as causation, but it is not "the same point".

So could we argue that the problem here really lies in the action at a distance? No, we are going to find that when action at a distance is ruled out on Lewis' theory, the reasons it is ruled out are exactly those reasons we've already seen: either the problems arise because we are dealing with simple and ideal cases, or because we find we cannot hold the right things fixed in the action at a distance region. And those problems can arise without action at a distance, cf Sections 3 and 4. We have already seen how this works for simple and ideal cases of simultaneous causation and for a limited set of cases of future action at a temporal distance: it works because one can achieve a replication miracle to reproduce the effect by a small miracle and thereby generate further perfect match. In the case of a large complex effect like the appearance of a person, we have seen that we do not get causation because it is not possible to hold fixed the events in the background of the cause, and in fact nothing in the time travel region can be held fixed by the similarity measure.

7. Conclusions

Contrary to his explicit advertisment, Lewis' semantics does not allow for backwards causation or time travel. Essentially, the reason is that when a cause c has effects in its own causal past, the closest $\sim c$ worlds will be worlds which do not hold fixed the background of c, hence in general, an earlier effect of c will not be absent from all the closest $\sim c$ worlds; $\sim c$ worlds which do hold fixed the background of c contain large miracles, and hence are not the closest worlds. The exception to this argument is the case where the time travel region and the causal past of a cause do not overlap, and where nothing in the time travel region causes anything in the causal past. The constraints necessary for this exception are difficult to instantiate. They do not in general allow for the kinds of physical hypotheses Lewis hoped to allow for, namely tachyons, advanced potentials, and cosmological models with closed timelike curves. These constraints may allow backwards causation in, for example, the case of a very specific trajectory of a tachyon, although they do not in general allow tachyon trajectories to count as backwards causation. Finally, these constraints certainly do not make possible the kinds of time travel Lewis has in mind—stories such as Tim attempting to kill his grandfather. This is significant not the least on account of the level of influence that Lewis' theory of causation and his defence of time travel have had on philosophy over the last 50 years.

Funding: This research was supported by the Australian Reasearch Councile grant #DP110101815.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 This argument was first presented at the American Philosophical Association Central Division meeting in Chicago, 18–21 February 2009.

2 Strictly speaking Lewis' similarity measure is a three place relation between two sets of worlds A, B say and some specific world, call it actual. It might give the result that each and every world in set A is closer to the actual world than is any world in set B. However, in rough and ready parlance, it is common practice to speak of the closest worlds.

3 The term 'lion' arose from consideration of Lewis-type time travel consistency claims: Tim attempts to kill his grandfather before his parents' conception, but fails for 'commonplace reasons', the gun jams, he loses his nerve etc. Do these reasons have their source in the region prior to the Cauchy horizon? Not necessarily: "an unexpected hungry lion behind the door of a time machine could effectively reconcile the traveler's freedom of will and his grandfather's safety" [14] (p. 064013–0640134); [15] (p. 183).

4 This is somewhat like the theory of Feynman [18] that positrons could be interpreted as electrons traveling backwards in time except that the turnaround at t_2 would require energy input which needs to be absent from our example.

References

1. Lewis, D. Causation. *J. Philos.* **1973**, *70*, 556–567. [CrossRef]
2. Lewis, D. Causation as Influence. *J. Philos.* **2000**, *97*, 182–198. [CrossRef]
3. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
4. Lewis, D. Counterfactual Dependence and Time's Arrow. *Noûs* **1979**, *13*, 455–476. [CrossRef]
5. Jackson, F. A Causal Theory of Counterfactuals. *Australas. J. Philos.* **1977**, *55*, 3–21. [CrossRef]
6. Fernandes, A. Time Travel and Counterfactual Asymmetry. *Synthese* **2019**, *198*, 1983–2001. [CrossRef]

7. Friedman, J.; Morris, M.; Novikov, I.; Echeverria, F.; Klinkhammer, G.; Thorne, K.; Yurtsever, U. Cauchy Problem in Spacetimes with Closed Timelike Curves. *Phys. Rev. D* **1990**, *42*, 1915–1930. [CrossRef] [PubMed]
8. Arntzenius, F.; Maudlin, T. Time Travel and Modern Physics. In *Time, Reality and Experience*; Callender, C., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 169–200.
9. Dowe, P. Constraints on Data in Worlds with Closed time-like Curves. *Philos. Sci.* **2007**, *74*, 724–735. [CrossRef]
10. Lewis, D. A Subjectivist's Guide to Objective Chance. In *Studies in Inductive Logic and Probability, Volume II*; Jeffrey, R., Ed.; University of California Press: Berkeley, CA, USA, 1980; pp. 263–293.
11. Dowe, P. Causal Loops and the Independence of Causal Facts. *Philos. Sci.* **2001**, *68*, S89–S97. [CrossRef]
12. Collins, J.; Hall, N.; Paul, L. Counterfactuals and Causation: History, Problems and Prospects. In *Causation and Counterfactuals*; Collins, J., Hall, N., Paul, L., Eds.; MIT Press: Cambridge, MA, USA, 2004; pp. 1–57.
13. Price, H. *Time's Arrow and Archimedes' Point*; Oxford University Press: New York, NY, USA, 1996.
14. Krasnikov, S. Time Travel Paradox. *Phys. Rev. D* **2002**, *65*, 064013. [CrossRef]
15. Krasnikov, S. *Back-in-Time and Faster-than-Light Travel in General Relativity*; Springer: Cham, Switzerland, 2018.
16. Wasserman, R. Lewis on Backward Causation. *Thought* **2015**, *4*, 141–150. [CrossRef]
17. Adlam, E. Spooky Action at a Temporal Distance. *Entropy* **2018**, *20*, 41. [CrossRef] [PubMed]
18. Feynman, R. The Theory of Positrons. *Phys. Rev.* **1949**, *76*, 749–759. [CrossRef]

philosophies

MDPI

Article

Lessons from Grandfather

Andrew Law and Ryan Wasserman *

Department of Philosophy, Western Washington University, Bellingham, WA 98225, USA; lawa@wwu.edu
* Correspondence: ryan.wasserman@wwu.edu

Abstract: Assume that, even with a time machine, Tim does not have the ability to travel to the past and kill Grandfather. Why would that be? And what are the implications for traditional debates about freedom? We argue that there are at least two satisfactory explanations for why Tim cannot kill Grandfather. First, if an agent's behavior at time *t* is causally dependent on fact *F*, then the agent cannot perform an action (at *t*) that would require *F* to have not obtained. Second, if an agent's behavior at time *t* is causally dependent on fact *F*, then the agent cannot perform an action (at *t*) that would prevent *F* from obtaining. These two explanations have distinct upshots for more traditional debates over freedom. The first implies that causal determinism is incompatible with the ability to do otherwise and also raises questions about the traditional arguments for the incompatibility of divine foreknowledge and the ability to do otherwise; the second does neither. However, both explanations imply that the Molinist account of divine providence renders agents unable to do otherwise, at least in certain circumstances.

Keywords: time travel; grandfather paradox; ability; freedom; fixity

Citation: Law, A.; Wasserman, R. Lessons from Grandfather. *Philosophies* **2022**, *7*, 11. https://doi.org/10.3390/philosophies7010011

Academic Editor: Alasdair Richmond

Received: 4 November 2021
Accepted: 19 January 2022
Published: 25 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Tim despises Grandfather and desires nothing more than to kill him. But alas, he is too late: Grandfather died of natural causes when Tim was only a child. Tim is not so easily deterred though. He builds a time machine, travels back to a time before his grandparents first met, and approaches Grandfather on the street, gun in hand, hatred in his heart.[1]

Of course, we know that Tim will not kill Grandfather, since part of the story is that Grandfather died of natural causes. But *can* he do so? That is, does Tim have the *power* or *ability* to kill Grandfather?[2]

Many people are inclined to think not. Indeed, this thought plays a key role in one of the most common arguments against the possibility of time travel: if time travel *were* possible, then a person like Tim *could* retroactively kill his grandfather. But Tim *cannot* do that, so time travel is impossible.

We will not defend the traditional answer here, though we believe that such a defense is possible.[3] Instead, we are interested in where this answer leads us. In particular, we will focus on two questions: (i) If Tim cannot kill Grandfather, then what is the best *explanation* for this fact? (ii) What, if anything, does this explanation teach us about other, more familiar debates about freedom? Regarding question (i), we will argue that there are at least two plausible explanations for why Tim cannot kill Grandfather. Regarding (ii), we will aim to show that both explanations render certain views about freedom problematic. But, before we get there, we will consider some less promising explanations for why Tim cannot kill Grandfather.

2. You Cannot Do the Impossible

The first explanation for Tim's inability is simple: Tim cannot kill Grandfather because doing so would be impossible.

If this was the correct explanation for Tim's inability, then the Grandfather paradox would have nothing new to teach us about the topic of freedom. After all, our inability to do the impossible is something that most of us already take for granted.[4]

However, the current explanation is *not* correct—or at least not complete. First, note that the act-type in question—"killing Grandfather"—is not in fact impossible. For example, many people could have killed Grandfather prior to Tim's arrival in the past. Indeed, if we assume that Grandfather lived long enough to overlap with Tim's childhood, then it was possible for *Tim* to kill Grandfather (in the traditional way, without traveling back in time). So, the explanation must be that it is impossible for *Tim* to kill Grandfather, and then only *at the relevant time*. However, once we make these changes, the explanation appears circular, since the claim that it is *impossible* for Tim to kill Grandfather (at the relevant time) seems equivalent to saying that Tim *cannot* (then) kill Grandfather.

In order to avoid the charge of circularity, one would need to provide an alternative understanding of "impossibility". For example, one could say that Tim is unable to kill Grandfather (at the relevant time) because his doing so would be *physically* impossible or *metaphysically* impossible or even *logically* impossible. On this way of looking at things, Tim cannot kill Grandfather (at the relevant time) for the same reason that he cannot accelerate past the speed of light, or drink water without H_2O, or prove *modus ponens* invalid—each of these things would be incompatible with the relevant laws (of physics, metaphysics, or logic).

This explanation would avoid the charge of circularity, but only at the cost of making implausible claims about physical, metaphysical, or logical possibility. After all, it does *not* seem impossible, in any of these senses, for Tim to kill Grandfather in the past. Consider a world where Grandfather miraculously rises from the dead in order to sire Father, who sires Tim, who travels back in time to kill Grandfather. Or consider a world where, right as Tim's bullet strikes his target, Grandfather fissions, leaving one living Grandfather and one dead (the living Grandfather going on to sire Tim's father, etc.). Or simply consider a world in which someone harvests Grandfather's genetic material shortly after his death and delivers it to Grandmother (so that she can sire Tim's father, etc.). These would all seem to be possible situations in which Tim kills Grandfather at the relevant time.[5] So, Tim's inability cannot be explained by the impossibility of that act.

Of course, one might point out that the imagined possibilities are irrelevant to the case at hand. If there are no genetic harvesters waiting in the wings, then the physical possibility of such beings is irrelevant to the question of whether or not Tim can kill Grandfather. The same thing is true for resurrection and fission since (we can assume) those things are not possible in Tim's situation. So, one might try to explain Tim's inability to kill Grandfather (at the relevant time) by saying that it is physically (or metaphysically, or logically) impossible for him to do so *given his actual circumstances*.[6] This explanation would not be obviously circular. Nor would it violate our intuitions about what is possible. But the challenge would be to say *which* features of Tim's circumstances are relevant and *how* they explain his inability to kill Grandfather. After all, to take a case inspired by David Lewis, it would be possible for a physical duplicate of Tim in physically identical surroundings to kill a physical duplicate of Grandfather [1]. So, the features of Tim's local environment cannot be what explains his inability. The question then remains: what is it about Tim's circumstances that render him unable to kill Grandfather? As we will see, there are several different ways of answering this question.

3. You Cannot Do Anything Incompatible with the Past

A second explanation for Tim's inability begins with the extremely intuitive idea that the past is "over and done with," "settled," or "fixed." Therefore, Tim's attempting to kill Grandfather in the past would be no more useful than his crying over spilt milk. Or more carefully, this explanation starts with the following principle:

Fixity of the Past (FP): For any agent, *S*, action *X*, and time *t*, if it is true that, were *S* to do *X* at *t*, some fact about the past (relative to *t*) would have been different, then *S* cannot do *X* at *t*.[7]

Now suppose that Tim is considering stepping into his time machine in the year 2021 to travel back and kill Grandfather in 1921. Since Grandfather in fact survived 1921, Tim's plan would require a fact about the past (relative to 2021) to be different. FP therefore says—correctly—that Tim cannot travel to the past and kill Grandfather.

Moreover, this explanation would seem to have significant implications for various debates over freedom. First, FP is often a crucial ingredient in arguments for the incompatibility of freedom and causal determinism. If causal determinism is true, then all of our actions are entailed (and caused) by the past and the laws. So, if we were to do otherwise, either the past or the laws would have to have been different. But, according to FP, we cannot perform an action that would require the past to be different. And it would seem as if we likewise cannot perform an action that would require the laws to be different. So, causal determinism is incompatible with the freedom to do otherwise.[8]

Second, FP plays a central role in arguments for the incompatibility of freedom and divine foreknowledge. If there is an infallible God who knows and has always known everything we will ever do, then, if we were to do otherwise, some of God's past beliefs would have to have been different than what they actually are. But according to FP, we cannot perform an action that would require the past to be different. So, divine foreknowledge is incompatible with the freedom to do otherwise.[9]

Both of these arguments have generated an *enormous* literature. However, we will not be addressing either argument (in its current form), since FP is *not* a satisfactory explanation of Tim's inability to kill Grandfather. There are two problems with this account, one obvious and one less so.

The obvious problem is this: while FP implies that Tim cannot, *in the year 2021*, travel to the past and kill Grandfather, it does not imply that he cannot, *in the year 1921*, carry out his plan. That's because the fact that Grandfather survived 1921 is only a fact about the past "before" Tim travels back to that time. Once he exits his time machine, Grandfather's survival is no longer in the past, so FP gives us no reason to think that Tim cannot carry out his plans.

The less obvious problem is this: while Tim, in the year 2021, cannot travel to the past and kill Grandfather, it does seem as if he can travel to the past and do something more mundane—he could, for example, kick over a rock in the year 1921, even if the rock was not in fact kicked over at that time. Or, at the very least and at first glance, the claim that Tim cannot kill Grandfather is far more intuitive than the claim that he cannot kick over the rock.[10] This, however, runs contrary to FP. The rock's not being kicked over in the year 1921 is just as much a part of the past, relative to 2021, as Grandfather's survival. That is, if Tim were to travel to the past and kick over the rock, a fact about the past (relative to the year 2021) would have to have been different. So, FP implies that, in the year 2021, Tim cannot even travel back and perform the simplest of actions (apart from those that he actually performs). This seems incorrect. Indeed, if time travel is possible, this constitutes a counterexample to FP.[11]

These two problems together show, at the very least, that Grandfather's survival being in the past is not enough to explain Tim's inability to kill him. But they also suggest a different way of understanding the claim that the past is fixed. This brings us to our third explanation.

4. You Cannot Do Anything Incompatible with Your Causal Past

Return to the difference between Tim's traveling to the past to kill Grandfather and his traveling to the past to kick over a rock. While there may not be a *temporal* difference between the two—both Grandfather's survival and the rock's not being kicked are in the past—there is an important *causal* difference. Grandfather's survival is a cause (or a cause of a cause . . .) of Tim's existence and, hence, of his traveling to the past in an attempt to

kill Grandfather; the rock's not being kicked over plays no such role. That is, while both Grandfather's survival and the rock's not being kicked over are part of the *temporal* past (relative to the year 2021), only the former is a part of Tim's *causal* past, where an agent's causal past includes all of those events that are causes (or causes of causes . . .) of the agent's present experiences and behavior. Given the foregoing shortcomings of FP, one might think that it is not the *temporal* past that is "settled" or "fixed" for agents, but the *causal* past instead. Under such a view, it is not the temporal relation per se between an agent and certain facts that matters for control; it is the causal relation instead.

One way to formalize this idea is as follows:

Fixity of the Causal Past (FCP): For any agent, S, action X, and time t, if it is true that, were S to do X at time t, some fact about S's causal past (relative to t) would have been different, then S cannot do X at t.[12]

For non-time travelers, such as ourselves, the difference between FP and FCP might be of little significance (although we will return to this shortly). But for individuals like Tim, the difference could not be more crucial. Grandfather's survival is part of Tim's causal past both before and after he enters his time machine. So, FCP implies that Tim lacks the ability to kill Grandfather at both of those times. However, while the rock's not being kicked over may be part of the temporal past (before Tim enters his time machine), it is not (necessarily) a part of his causal history.[13] So, FCP does not imply that he cannot kick over the rock. FCP is thus a more promising explanation of Tim's inability to kill Grandfather.[14]

FCP would also have significant implications for many debates over freedom. For instance, if causal determinism is true, then our causal pasts necessitate all of our behavior. Since FCP claims that our causal pasts are fixed in determining what we are free to do, FCP implies that causal determinism is incompatible with our being free to do otherwise. Or return to the issue of divine foreknowledge. Even if God's infallible beliefs are part of the *temporal* past, they are (arguably) not part of our *causal* pasts, in which case FCP does not imply that divine foreknowledge is incompatible with our being free to do otherwise. This would be a significant result, for it would imply that determinism undermines freedom in way that foreknowledge does not.[15]

While FCP is an improvement over FP, it still has a significant shortcoming, one involving a slightly different asymmetry. Suppose that Tim travels to the past and visits himself as an infant. Struck by his own cuteness, Adult Tim considers pinching Infant Tim on the cheek but ultimately decides against it. Intuitively, even though Tim does not pinch his infant self (or even attempt to), he could have. Or, at the very least, it is not obvious that he is unable to do so.

Once we admit of this asymmetry in Tim's abilities, FCP looks problematic. After all, the fact that Infant Tim was not pinched at the time in question would be a fact about Tim's causal past since it would directly concern his own earlier experiences. So, FCP implies that Tim *cannot* pinch his earlier self, any more than he can kill Grandfather. This seems incorrect.

Fortunately, there is a relatively simple way of getting around this problem.

5. You Cannot Do Anything Incompatible with the Causal Past of Your Behavior

The simple move is this: instead of holding fixed the *agent's* causal past in determining what they can do, we ought to hold fixed the causal past of the agent's *behavior*. Return to the difference between Tim's not killing Grandfather and his not pinching his infant self. The absence of each of these events may be part of *Tim's* causal past, but only the former absence is a cause (or a cause of a cause . . .) of Tim's traveling to the past and his other subsequent actions. Infant Tim's not being pinched might be a cause of *some* of Tim's behavior (e.g., it might be a cause of his not crying at that moment), but it does not play a causal role in his decision to travel back in time or his decision to refrain from pinching his younger self.

This suggests the following revision of FCP:

Fixity of the Causal Past—Behavior (FCP-B): For any agent *S*, action *X*, and time *t*, if it is true that, were *S* to do *X* at *t*, some fact about the causal past of *S*'s *behavior* at *t* would have been different, then *S* cannot do *X* at *t*.

FCP-B respects the intuitive verdict in all of our cases so far. It implies that Tim cannot kill Grandfather, but it does not imply that Tim cannot kick over the rock or that he cannot pinch his earlier self. Moreover, it still delivers significant results for the debates over freedom: for similar reasons as those rehearsed above, it also seems to imply that causal determinism, but not necessarily divine foreknowledge, is incompatible with our being free to do otherwise.

However, there is one final problem to address. Consider a case of overdetermination. Suppose that, before stepping into his time machine, Tim draws up two qualitatively identical sets of plans for his time machine. He then travels back in time, tracks down his younger self, and passes off both sets of plans. Young Tim sees both sets simultaneously and, over time, uses that information to build a time machine. He then draws up two qualitatively identical sets of plans, enters into the time machine, and sets off to find his younger self . . .

Now add a small wrinkle to the story: suppose that, just before Old Tim gave Young Tim both sets of plans, he considers destroying one set. (Perhaps he is worried that the second set could fall into the wrong hands.) Old Tim eventually decides against destroying either set, and passes on both sets of plans to Young Tim.

While Old Tim did not destroy a set of plans, it nonetheless seems as if he could have. He could not have destroyed *both* sets, of course, since he would then never be in a position to travel back in time with the plans to begin with. But destroying *one* set seems harmless, since each set is causally sufficient for Tim's actual journey.[16]

FCP-B implies the opposite. Since each set of plans is a cause of Tim's behavior (including his trip to the past), FCP-B implies that he cannot destroy either set. In order to avoid this result, we will have to make one final adjustment. Instead of holding fixed *every* cause of an agent's behavior, we should only hold fixed those causes on which the agent's behavior depends. Here, *x* is *causally dependent* on *y* just in case (i) *x* would not have occurred (or been the case) if *y* had not occurred, and (ii) this counterfactual holds at least in part because *y* is a cause of *x*. In the case of overdetermination, the survival of each set of plans is a cause of Tim's journey to the past, but his journey to the past does not counterfactually depend on either set. If Tim were to destroy the second set, his younger self would still receive the first set, which is all that is required for Tim's journey to the past (as well as his writing down both sets of plans). So, by definition, his journey to the past does not causally depend on either set.

To capture this thought, we can (one last time!) revise our principle as follows:

*Fixity of the Causal Past—Behavior** (FCP-B*): For any agent *S*, action *X*, and time *t*, if it is true that, were *S* to do *X* at *t*, some fact that *S*'s behavior at time *t* is (actually) causally dependent on would not have occurred, then *S* cannot do *X* at *t*.

We have finally arrived at a satisfactory explanation of Tim's inability to kill Grandfather. Like its predecessors, this principle has significant implications for freedom. Again, FCP-B* would seem to imply that causal determinism is incompatible with our being free to do otherwise, at least in those cases where our behavior is not causally overdetermined. Moreover, just as with FCP and FCP-B, our final principle does *not* imply that divine foreknowledge is incompatible with our being free to do otherwise. It is true that our behavior would *counterfactually* depend on God's past beliefs. But counterfactual dependence and causal dependence are two different things. Assuming that our behavior is not caused by God's past beliefs, FCP-B* does not apply to the case of divine foreknowledge.

These would be fascinating lessons from Grandfather. Unfortunately (or perhaps fortunately), concluding that FCP-B* is the correct explanation is too hasty. That's because there is a simpler and more restricted principle that can do the same work.[17]

6. You Cannot Do Anything That Would Be Self-Undermining

One common reason for thinking that Tim cannot retroactively kill Grandfather is that such an act would be, in some sense, *self-defeating*.[18] After all, if Tim were to kill Grandfather, then Father would never be born, in which case *Tim* would never be born, in which case Tim would *not* go back to kill Grandfather. So, Tim's murder of Grandfather would be "self-undermining" in the sense that, were he to do it, Tim would not be there to kill Grandfather in the first place. Plausibly, no agent can perform an action of this kind, which would explain why Tim is unable to kill Grandfather.

There is something appealing about this rationale, but it needs to be made more precise. The rough characterization of a "self-undermining" act just given cannot be right. According to that characterization, Tim's killing Grandfather is a self-undermining act if the closest world where he kills Grandfather is a world in which he does not. But there are no (possible) worlds like that, so the counterfactual—*if Tim were to kill Grandfather, then he would not*—is trivially true, at best. At worst, the counterfactual is false since, as already noted, there seem to be (possible) worlds where Tim successfully kills Grandfather. Without a more plausible characterization of a "self-undermining" act, this explanation is incomplete.

Fortunately, our reflections on causation, behavior, and dependence point toward a better characterization. We have already noted that Tim's attempt to kill Grandfather is causally dependent on Grandfather's survival. So, Tim's attempt to kill Grandfather is an attempt to perform an action such that, were he to successfully perform it, some event that his attempt is causally dependent on would not have occurred. But Tim's attempt to kill Grandfather is more than that: his attempt, if successful, would *prevent* an event that his very attempt is causally dependent on. That is, there's not only a *counterfactual* relation between the success of his attempt and his attempt's causal history (namely, if he were to succeed, then an event that his attempt is causally dependent on would not have occurred), but a *causal* relation as well (namely, if he were to succeed, then he would *prevent* such an event).

With this in mind, one could hold that there is no difficulty in an agent performing an action such that, were she to perform it, some event that her behavior is causally dependent on would *not* have occurred. The difficulty only arises if the agent's action would *cause* the event to not occur. Or, more formally:

No Self-undermining Actions (NSA): For any agent *S*, action *X*, and time *t*, if it is true that, were *S* to do *X* at time *t*, *S* would thereby prevent some fact that *S*'s behavior at time *t* is (actually) causally dependent on, then *S* cannot do *X* at *t*.

Like FCP-B*, NSA respects the intuitive verdicts in our previous cases. If Tim were to kick over the rock, pinch his younger self, or destroy one set of plans, he would not thereby *prevent* some fact that his (actual) behavior is causally dependent on. Thus, NSA does not imply that Tim is unable to do those things.

However, unlike FCP-B*, NSA does *not* imply that causal determinism is incompatible with freedom. To illustrate, suppose that causal determinism is true and that Tim picks up his coffee cup at time t. If Tim had *not* picked up the cup at that time, then (given determinism) some event that his actual behavior causally depends on would have to have been different—perhaps a certain neuron that fired just prior to his picking up the cup would not have fired in that counterfactual scenario.[19] But his refraining from picking up the coffee cup in that case would not have *prevented* the earlier firing of the neuron. More generally, doing otherwise in a deterministic world requires a certain counterfactual connection between our behavior and the events that our behavior is causally dependent on. But it does not require that, were we to do otherwise, we would thereby *cause* some of those events to be different. So, doing otherwise in a deterministic world does not violate NSA.

What about the issue of divine foreknowledge and freedom? Like FCP-B*, NSA does not imply that divine foreknowledge and freedom are incompatible. If we were to do otherwise than what we actually do, God would have had a different set of beliefs in the past. But, as already noted, our actual behavior is not *causally* dependent on God's past

beliefs. Moreover, it is at least arguable that our doing otherwise would not have *prevented* God from holding his actual, earlier beliefs.[20]

While NSA fails to imply that freedom is incompatible with causal determinism or divine foreknowledge, it is less clear how significant this is. As we saw above, the principle of the Fixity of the Past (FP) is a driving force behind the standard arguments that causal determinism and divine foreknowledge are both incompatible with freedom. Even if one rejects FP (as we think one ought to), there is still some pressure to account for whatever it is that we find intuitive about this principle. For instance, suppose that Tim leaves his house very late and, thus, is sure to be tardy for the faculty meeting. Is he now able to arrive on time? Presumably not. But if we reject FP, then what accounts for Tim's inability in this case? Without an account, there is always the danger that whatever kernel of truth there is in FP will be enough to establish the incompatibility of freedom with causal determinism and foreknowledge.

This is one of the reasons principles like FCP and its variations are so significant: in addition to delivering intuitive results in various time travel cases (and not delivering unintuitive ones), they seem to capture the appeal of FP in ordinary contexts. If Tim were to arrive on time to the faculty meeting, it is not just that some fact about the temporal past would have been different, but, arguably, some fact about Tim's causal past, or the causal history of his behavior, would have been different as well. (Either he would not have left so late, or there would not have been so much traffic, etc.) That means FCP and its variations might be suitable *replacements* for FP—they may capture the intuitive idea behind that principle.

This is in stark contrast with the relation between NSA and FP. Whatever intuitive appeal FP has, it is *not* captured by NSA. If Tim were to arrive on time to the faculty meeting, he would not thereby *prevent* some fact that his actual behavior is causally dependent on, and so NSA does not imply that Tim cannot arrive on time. Any suitable replacement for FP ought to imply as much. Hence, those who are sympathetic to the standard arguments for the incompatibility of freedom and causal determinism or divine foreknowledge should not feel threatened by the truth of NSA whatsoever. Or another way to put the point: while there would seem to be little motivation for accepting *both* FCP (or some variation thereof) and FP, there is plenty of motivation for accepting both NSA and FP. If so, then the fact that NSA fails to imply that freedom is incompatible with causal determinism or foreknowledge is of little to no significance. NSA is simply orthogonal to such debates.

This might leave one wondering whether there are *any* significant implications of NSA. We believe that there is at least one, and it involves another theological example. "Molinism" or "Middle-knowledge theory" is a popular account of how God exercises providential control over the world (assuming that God exists). Roughly, it is comprised of (at least) three claims: (i) for any possible agent, S, and any possible circumstance, C, that S might be in, there is a fact of the matter as to what S would freely do in C; (ii) such facts, often called "counterfactuals of creaturely freedom," are contingent and independent of God's will; and (iii) God uses knowledge of such facts in determining which world to actualize. As a simplistic example, suppose God is deciding whom to place in the garden. Looking over all the possible creatures, and knowing what each creature would freely do, were they placed in the garden, God ranks Adam as the best overall choice with Ben a very close second. However, Adam is only ranked ahead of Ben because Adam would freely eat of the fruit, were he placed in the garden; if it were not the case that Adam would freely do so, then God would have ranked Ben over Adam. So, God places Adam in the garden and Adam freely eats of the fruit, God of course knowing that he would do so.

Is Adam able to refrain from eating the fruit? Not if NSA (or FCP-B*) is correct. Presumably, Adam's behavior in the garden is causally dependent both on God's decision to place him in the garden (or actualize a world where Adam is in the garden) and God's belief that Adam would freely eat the fruit, were Adam placed in the garden. So, according to FCP-B*, both God's decision and God's belief are fixed in assessing whether Adam could refrain from eating the fruit, which means that Adam cannot refrain. The same result

holds under NSA. If Adam were to refrain from eating the fruit, he would have thereby prevented God from believing that, were Adam placed in the garden, he would freely eat the fruit (since that counterfactual of creaturely freedom would not be true). So, if Adam were to refrain from eating the fruit, he would thereby prevent a fact (or event) that his actual behavior is causally dependent on. So, NSA implies that Adam cannot refrain.[21]

What should the Molinist make of this? One option is to reject the starting assumption of this paper and insist that Tim can kill Grandfather. One could then reject our case for FCP-B* and NSA and insist that Adam could have refrained from eating the fruit (despite the fact that this would be a self-undermining act). In our view, a more promising option is to instead insist that an agent can freely perform an action even if the agent is not able to do otherwise—that Adam freely eats of the fruit even though he could not have refrained from doing so.[22] Admittedly, the view that freedom does not require the ability to do otherwise is not a novel one, even in the context of Molinism[23] or time travel.[24] But what is noteworthy is that, if our starting assumption that Tim cannot kill Grandfather is correct, and if either FCP-B* or NSA is the correct explanation for that assumption, then Molinists *must* accept this conclusion. Many will deem that a cost for the Molinist view of providence.

7. Conclusions

Where has our starting assumption led us? First, if we are correct, then the right way to think about the fixity of the past is in terms of causation, rather than time. Given the prominent role the fixity of the past plays in many contemporary arguments regarding freedom, this is an important result. Second, regardless of whether FCP-B* or NSA is the correct explanation for Tim's inability to kill Grandfather, it appears that Molinist views of providential control limit the abilities of agents in interesting ways. Both of these claims, we suggest, are important lessons to take away from Grandfather.[25]

Author Contributions: Conceptualization, A.L. and R.W.; methodology, A.L. and R.W.; writing—original draft preparation, A.L. and R.W.; writing—review and editing, A.L. and R.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

1 We borrow this familiar story from David Lewis. See [1].

2 "Can" is notoriously context sensitive and does not always concern agential abilities. If murder were made legal, then Tim "could" kill Grandfather in the sense that his doing so would be compatible with the relevant laws. But this would not mean that Tim had the *ability* to kill Grandfather, any more than overturning an ordinance aginst transubstantiation would grant him the power to turn water into wine. Our focus throughout will be on abilities.

3 We here have in mind the kind of argument given by Kadri Vihvelin—see [2,3] for discussion. For potential replies, see [4,5]. Vihvelin's argument famously relies on the principle that "if someone would fail to do something, no matter how hard or how many times she tried, then she cannot do it." While we would endorse a version of this principle, it leaves the more basic question unanswered: *why* would Tim fail to kill Grandfather, no matter what he tried? Without answer to this question, we would not have an explanation of Tim's inability. (For one possible answer, see Chapter 4, Section 3 of [3].)

4 For some exceptions to this rule, see [6,7].

5 See [2,3] and, especially, [8] for more cases of this kind.

6 Thank you to Dan Howard-Snyder and Neal Tognazzini for pressing this point.

7 See [9] for the classic formulation, but also [10–12].

8 For a more detailed treatment of this argument, see [9].

9 For the classic introduction to this argument, see [13].

10 We are assuming that the rock's being kicked over need not affect Tim's life up until 2021 in any significant way—that there is not necessarily a "butterfly effect" in this case. This is not to say that the rock could not have had such an effect. But, if it did, then our intuitions would shift. In particular, if Tim's trip to the past causally depended on the rock *not* being kicked, then it would

seem as if Tim *cannot* kick over the rock. For simplicity, we will assume that there are no "butterfly effects" in the following cases as well.

11 For more on the problems with FP, see [14–17].

12 Michael Rea employs a similar principle in [18]. See also [15,17].

13 For a counter to this claim, see [18]. Again, for the sake of simplicity, we will assume that there is no causal connection of this kind.

14 FCP would also respect Lewis's famous asymmetry between Tim and Tom [1] (p. 149). For discussion, see [19].

15 See [15,17] for more on the relationship bewteen FCP and theological fatalism. On the more general point that causal determinism seems to threaten freedom in a way that foreknowledge does not, see [20] (p. 412), [21] (p. 89), and, especially, [22] (p. 1012).

16 One of the authors would like to thank Justin Mooney for helping develop this point.

17 To be clear, we are not denying FCP-B*. On the contrary, at least one of the authors of this paper accepts FCP-B*. Rather, the point is that one need not accept FCP-B* to account for Tim's inability to kill Grandfather, since the principle in the next section looks equally promising (at least as an explanation of Tim's inability to kill Grandfather). Thank you to Dennis Whitcomb for suggesting this clarification.

18 See, e.g., [1] (p. 152), [2] (p. 315), and [3] (p. 72).

19 Here, we continue to assume the standard view that, given determinism, the closest worlds where we act otherwise are ones where both the laws *and* the past are different. We also set aside the issue of causal overdetermination.

20 However, see Section 5 of [23], where it is suggested that some of God's past beliefs are causally dependent on our future actions.

21 We are assuming that God's decision to actualize a world where Adam is in the garden is a *cause* (or a cause of a cause . . .) of Adam's eating the fruit, and also that the relevant counterfactual of creaturely freedom is a cause of God's creative decision. Both of these claims could be challenged. Nonetheless, there is clearly a sense in which God's decision to actualize the relevant world *explains* (at least ancestrally) Adam's eating the fruit. So too, there is a natural sense in which God's creative decision is explained (in part) by the relevant counterfactual of creaturely freedom. And it would not be difficult to amend FCP-B* or NSA to focus on *explanatory dependence* rather than *causal dependence*. For simplicity, we will ignore these complications.

22 Molinists appear to be split on whether freedom requires the ability to do otherwise. Molinism's namesake, Luis de Molina, seems to accept that claim, writing: "the third type is middle knowledge, by which, in virtue of the most profound and inscrutable comprehension of each faculty of free choice, [God] saw in His own essence what each such faculty would do with its innate freedom were it to be placed in this or in that or, indeed, in infinitely many orders of things—even though it would really be able, if it so willed, to do the opposite . . . " [24] (p. 52). However, some contemporary Molinists, such as William Craig [25], deny this claim.

23 The issue of whether freedom (in the sense required for moral responsibility) requires the ability to do otherwise has received a vast amount of attention ever since Harry Frankfurt's pivotal piece [26], and even in the context of Molinism. (See the back-and-forth between [27–29].)

24 Several authors have used cases of time travel to motivate this thought. See [30,31] for two examples.

25 For helpful discussion and feedback, we thank Taylor Cyr, Dan Howard-Snyder, Frances Howard-Snyder, Hud Hudson, Christian Lee, Justin Mooney, Neal Tognazzini, and Dennis Whitcomb.

References

1. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Quarterly* **1976**, *13*, 145–152.
2. Vihvelin, K. What time travelers cannot do. *Philos. Stud.* **1996**, *81*, 315–330. [CrossRef]
3. Wasserman, R. *Paradoxes of Time Travel*; Oxford University Press: Oxford, UK, 2018.
4. Kiourti, I. Killing Baby Suzy. *Philosophia* **2008**, *139*, 343–352. [CrossRef]
5. Hanley, R. Autoinfanticide Is No Biggie: The Reinstatement Reply to Vihvelin. *Philosophies* **2021**, *6*, 87. [CrossRef]
6. Spencer, J. Able to do the Impossible. *Mind* **2017**, *126*, 466–497. [CrossRef]
7. Effingham, N. *Time Travel: Probability and Impossibility*; Oxford University Press: Oxford, UK, 2020.
8. Carrol, J. Ways to Commit Autoinfanticide. *J. Am. Philos. Associat.* **2016**, *2*, 80–91. [CrossRef]
9. Fischer, J.M. *The Metaphysics of Free Will*; Blackwell: Oxford, UK, 1994.
10. van Inwagen, P. *An Essay on Free Will*; Oxford University Press: Oxford, UK, 1983.
11. Ginet, C. *On Action*; Cambridge University Press: Cambridge, UK, 1990.
12. Fischer, J.M. Freedom, Foreknowledge, and the Fixity of the Past. *Philosophia* **2011**, *39*, 461–474. [CrossRef]
13. Pike, N. Divine Omniscience and Voluntary Action. *J. Philos.* **1965**, *74*, 27–46. [CrossRef]
14. Swenson, P. Ability, Foreknowledge, and Explanatory Dependence. *Australas. J. Philos.* **2016**, *94*, 658–671. [CrossRef]
15. Law, A. The Dependence Response and Explanatory Loops. *Faith Philos.* **2020**, *37*, 294–307. [CrossRef]
16. Law, A. From the Fixity of the Past to the Fixity of the Independent. *Philos. Stud.* **2021**, *178*, 1301–1314. [CrossRef]
17. Wasserman, R. The Independence Solution to the Problem of Theological Fatalism. *Philos. Phenomenol. Res.* **2020**. [CrossRef]
18. Rea, M. Time Travelers are not Free. *J. Philos.* **2015**, *112*, 266–279. [CrossRef]
19. Wasserman, R. Time Travel, Ability, and Arguments by Analogy. *Thought* **2017**, *6*, 17–23. [CrossRef]

20. Fischer, J.M.; Tognazzini, N.A. The Physiognomy of Responsibility. *Philos. Phenomenol. Res.* **2011**, *82*, 381–417. [CrossRef]
21. Byerly, T.R. Infallible Divine Foreknowledge Cannot Uniquely Threaten Human Freedom, but its Mechanics Might. *Eur. J. Philos. Relig.* **2012**, *4*, 73–94. [CrossRef]
22. Law, A.; Tognazzini, N.A. Free Will and Two Local Determinisms. *Erkenntnis* **2019**, *84*, 1011–1023. [CrossRef]
23. Wasserman, R. Freedom, Foreknowledge, and Dependence. *Noûs* **2021**, *55*, 603–622. [CrossRef]
24. Molina, L. *On Divine Foreknowledge: Part IV of the Concordia*; Freddoso, F., Ed.; Cornell University Press: Ithaca, NY, USA, 1988.
25. Craig, W.L. Response to Gregory A. Boyd. In *Four Views on Divine Providence*; Jowers, D., Ed.; Zondervan: Grand Rapids, MI, USA, 2011; pp. 224–230.
26. Frankfurt, H. Alternate Possibilities and Moral Responsibility. *J. Philos.* **1969**, *66*, 829–839. [CrossRef]
27. Bergmann, M. Molinist Frankfurt-style Counterexamples and the Free Will Defense. *Faith Philos.* **2002**, *19*, 462–478. [CrossRef]
28. Flint, T. On Behalf of the PAP-ists: A Reply to Bergmann. *Faith Philos.* **2002**, *19*, 479–484. [CrossRef]
29. Bergmann, M. Agent Causation and Responsibility: A Reply to Flint. *Faith Philos.* **2003**, *20*, 229–235. [CrossRef]
30. Spencer, J. What Time Travelers Cannot Do (But Are Responsible for Anyways). *Philos. Stud.* **2013**, *166*, 149–162. [CrossRef]
31. Tognazzini, N.A. Free Will and Time Travel. In *The Routledge Companion to Free Will*; Timpe, K., Griffith, M., Levy, N., Eds.; Routledge: New York, NY, USA, 2016; pp. 658–671.

philosophies

MDPI

Article

Back to the Present: How Not to Use Counterfactuals to Explain Causal Asymmetry

Alison Fernandes

Department of Philosophy, Trinity College Dublin, D02 PN40 Dublin, Ireland; asfernan@tcd.ie

Abstract: A plausible thought is that we should evaluate counterfactuals in the actual world by holding the present 'fixed'; the state of the counterfactual world at the time of the antecedent, outside the area of the antecedent, is required to match that of the actual world. When used to evaluate counterfactuals in the actual world, this requirement may produce reasonable results. However, the requirement is deeply problematic when used in the context of explaining causal asymmetry (why causes come before their effects). The requirement plays a crucial role in certain statistical mechanical explanations of the temporal asymmetry of causation. I will use a case of backwards time travel to show how the requirement enforces certain features of counterfactual structure *a priori*. For this reason, the requirement cannot be part of a completely general method of evaluating counterfactuals. More importantly, the way the requirement enforces features of counterfactual structure prevents counterfactual structure being derived from more fundamental physical structure—as explanations of causal asymmetry demand. Therefore, the requirement cannot be used when explaining causal asymmetry. To explain causal asymmetry, we need more temporally neutral methods for evaluating counterfactuals—those that produce the right results in cases involving backwards time travel, as well as in the actual world.

Keywords: causal asymmetry; time travel; counterfactuals; present; loewer

Citation: Fernandes, A. Back to the Present: How Not to Use Counterfactuals to Explain Causal Asymmetry. *Philosophies* **2022**, *7*, 43. https://doi.org/10.3390/philosophies7020043

Academic Editor: Alasdair Richmond

Received: 18 February 2022
Accepted: 28 March 2022
Published: 9 April 2022

1. Introduction

A plausible thought is that we should evaluate counterfactuals in the actual world by holding the present 'fixed'. More precisely, some methods require the state of the counterfactual world at the time of the antecedent to match that of the actual world, outside the spatial area of the antecedent, when evaluating counterfactuals in the actual world—where this requirement is explicitly part of the recipe for evaluating counterfactuals [1–3]. There might also be some changes to the present required to satisfy the content of the antecedent, perhaps implied by context [3] (p. 26). Being in Uganda may, in ordinary contexts, imply that you are not in Spain. However, other than changes within the spatial area of the antecedent or directly required to satisfy the antecedent, the state of the universe at the time of the antecedent remains unchanged.

For example, say you wonder what would be the case were you to frolic on the university lawns at midnight (given that you do not in the actual world). According to 'altered states recipe' approaches [1–3], the relevant nearby counterfactual world that determines what counterfactuals are true is one in which either your bodily movements or location at midnight are *different* from what they are in the actual world. Your frolicking on the lawns at midnight may imply that you are not resting on the lawn or not in the lounge. However, the states, at midnight, of the lawns, the university buildings, the security guards, and the entire rest of the universe outside the spatial areas you occupy in the counterfactual world (and in the actual world) are *required* to be the same as what they are in the actual world. While future states outside the spatial area of the antecedent (and that do not directly concern the antecedent) may differ, depending on the rest of the machinery for evaluating counterfactuals, present states may not. Thus, any counterfactual of the form 'If you were

to frolic on the university lawns at midnight, state x at midnight would obtain' (where x is different from what is the case in the actual world and does not occupy the spatial area of the antecedent or directly concern the content of the antecedent) is straightforwardly false.

Call an explicit requirement of this kind 'holding the distant present fixed'. This requirement looks reasonable. After all, evaluating counterfactuals seems to be about making minimal changes to actuality, and holding the distant present fixed seems to ensure minimal changes. However, while the requirement may produce reasonable results when evaluating counterfactuals in the actual world, it produces the wrong results in cases involving backwards time travel. The requirement implies strange counterfactual worlds, where there are changes to the past, but they are always 'put back' by the time of the present, by whatever means necessary. Assuming backwards time travel is possible, and such consequences unreasonable, holding the distant present fixed cannot be part of a completely general method of evaluating counterfactuals.

This result might seem trivial or uninteresting. After all, many approaches to evaluating counterfactuals were not designed for cases of backwards time travel. Indeed, defenders are sometimes explicit that their methods are not intended to cover cases involving backwards causation ([1], pp. 10–12; [2], p. 8) or only aim to recover counterfactuals true in the actual world [3] (pp. 21, 31). Thus, it is no wonder they produce odd results and no mark against them. Indeed, as far as I am aware, no one has even explicitly defended the idea that we should evaluate counterfactuals in backwards time travel scenarios by holding the distant present fixed. So, why does it matter if such approaches fail in these settings? Moreover, the project of determining the right way to evaluate counterfactuals in cases of backwards time travel might seem irrelevant to how we evaluate counterfactuals in the actual world. What could we possibly learn about evaluating counterfactuals in the actual world by considering time travel scenarios?

However, it turns out there is a project that relies on using a general method of evaluating counterfactuals—one that works in both backwards time travel scenarios and in the actual world. This is the project of using counterfactuals to explain temporal asymmetries in the actual world, particularly the temporal asymmetry of causation—why causes come before their effects at our world, hereafter 'causal asymmetry'. What motivates these accounts is less an *a priori* commitment to reducing causation, but the need to reconcile the temporal asymmetry of causation with temporal symmetry in the fundamental physical laws—see [4,5]. These accounts aim to trace causal asymmetry back to more fundamental physical asymmetries, using counterfactuals as a half-way step.

As I will argue (Section 3), the project of explaining causal asymmetry requires a temporally neutral method of evaluating counterfactuals—one that does not, *a priori*, enforce temporal features of counterfactual structure, but instead allows that structure to be derived from contingent physical structure in the universe. However, in attempting to explain causal asymmetry, certain statistical mechanical accounts use methods of evaluating counterfactuals that hold the distant present fixed and rely crucially on this requirement [6–8]. However, because the requirement to hold the distant present fixed enforces features of the counterfactual structure *a priori*, it prevents these features of counterfactual structure being derived from physical structure. More particularly, the requirement to hold the distant present fixed presumes there is no counterfactual dependence of the distant present. This presumption not only produces unreasonable results in cases where we expect such dependence, including backwards time travel cases, but it prevents what counterfactual dependencies there are from being derived from physical structure, in the way explanations of causal asymmetry demand. Therefore, the purported explanations do not adequately trace causal asymmetry back to more fundamental physical asymmetries, and so do not explain causal asymmetry.

Altogether, my target here is not 'altered states recipe' approaches that hold the distant present fixed when evaluating counterfactuals in the actual world. My target is those that have adopted methods that hold the distant present fixed when explaining causal asymmetry in the actual world.

The paper proceeds as follows. In Section 2, I argue that the requirement to hold the distant present fixed implies unreasonable results in a scenario involving backwards time travel. This outcome may seem obvious, given that the requirement was not designed to be used in cases of backwards time travel. However, understanding why the requirement produces unreasonable results in time travel scenarios allows us to see why the requirement is illicit when used to explain causal asymmetry. In Section 3, I argue that the failure of the requirement in time travel scenarios not only shows that the requirement cannot be part of a completely general method of evaluating counterfactuals, but more importantly, the failure of the requirement compromises certain statistical mechanical explanations of causal asymmetry.

One may have other reasons for rejecting the requirement to hold the present fixed, such as wanting to allow for simultaneous direct causation ([1], p. 9; [9], p. 426), backtracking counterfactuals [10] (p. 340), or other contextual requirements concerning 'how the change is to be effected' [3] (p. 26). The time travel counterexample I give does not rely on these concerns. I adopt a broadly Lewisian approach to time travel [11]: I assume that backwards time travel is possible and requires backwards causation. While the counterexample could be presented using merely backwards causation (without backwards time travel), time travel allows the point to be put particularly vividly—particularly when agents and decision-making are introduced (Section 3). For defences of the possibility of backwards time travel, see [11–13]. For other attempts to use backwards time travel scenarios to argue against certain methods of evaluating counterfactuals and chances, see ([14], chapter 7; [15–18]. I adopt a broadly B-theoretic view of time, such that talk of the past, present, and future is to be treated indexically—these are times before, simultaneous with, or after a given reference frame (usually the time of the counterfactual antecedent).

2. Time-Travelling Tina

Here is a case involving backwards time travel. I will use the case to demonstrate how evaluating counterfactuals by holding the present fixed enforces features of counterfactual structure, leading to unreasonable results in cases of backwards time travel. In the next section, I use this examination to argue that the requirement cannot be part of a suitably general method of evaluating counterfactuals that could be used to explain causal asymmetry.

Tina at the Beach: Tina is working from home. It is the end of a hot day, and, while she is sorry to have stayed in, she comforts herself with the following thought: in just a moment she will jump into her time machine, travel back in time to the start of the day, and spend the same day at the beach. "No doubt I'm there right now," she thinks, "feeling the warm sun and hearing the waves lapping". Tina is deliberating, however, about whether to shave her beard before going. [1] Assume Tina has no knowledge of her state at the beach and that her state at the beach is not a cause or causal condition of the decision she makes now—Tina's decision and state at the beach are not parts of a causal loop.[2]

Stipulate that Tina's time machine works, so that she will travel in time (and space) from her home (at time t_2) to the start of the day at the beach (time t_1) by reliable nomic and causal means, as represented in Figure 1. In Figure 1, 'Tina-at-home', 'Tina-arrived-at-beach', and 'Tina-tanned' refer to Tina at different spatiotemporal locations (assuming no particular metaphysics of persistence). Tina-at-home and Tina-tanned occupy different spatial locations, but the same temporal location. Tina's time travel may occur via curves in spacetime or other physical means. For simplicity, assume that Tina does not age significantly when she travels.

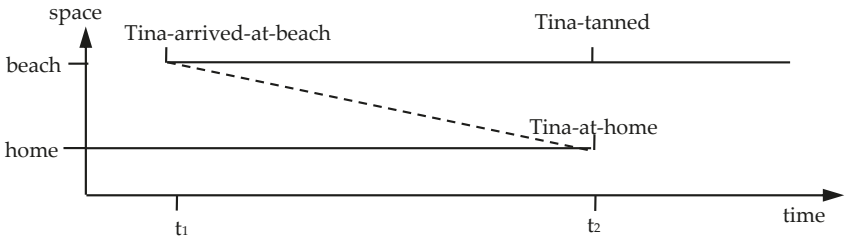


Figure 1. Tina's spatial location as a function of time. The dashed line indicates a form of time travel (whether continuous or discontinuous).

Consider the following counterfactuals:

1. If Tina-at-home were to shave, Tina-arrived-at-beach would be beardless.
2. If Tina-at-home were not to shave, Tina-arrived-at-beach would be bearded.

I claim these counterfactuals are true. First, they capture ordinary lawful and causal behaviour, where shaving at one time leads to be being beardless at later times, in a temporal order that aligns with causal order in the vicinity of the agent. There is no reason to expect violations of this ordinary behaviour, such as to avoid self-defeating causal loops—Tina's state at the beach is not a cause or causal condition of the decision she makes now. Second, 1 and 2 are the right counterfactuals for decision-making. Tina should decide whether to shave partly in light of whether she'd prefer to be bearded or beardless at the beach. Third, according to counterfactual accounts of causation, there must be counterfactual dependence of some of Tina's temporally earlier states on some of Tina's temporally later states for this to be a case of backwards time travel [11]. Thus, some counterfactuals of the form of 1 and 2 must be true.

Consider next the following counterfactuals:

3. If Tina-at-home were to shave, Tina-tanned would be beardless.
4. If Tina-at-home were not to shave, Tina-tanned would be bearded.

I claim these counterfactuals are also true. First, they capture ordinary lawful and causal behaviour, where shaving at one time leads to be being beardless at later times, in a temporal order that aligns with causal order in the vicinity of the agent. For the same reasons as above, there is no reason to expect violations to avoid self-defeating causal loops. Second, they are the right counterfactuals for decision-making: Tina should decide whether to shave partly in light of whether she would prefer to be bearded or not at the beach. Third, while there need not strictly be counterfactual dependence of Tina-tanned on Tina-at-home for this to be a case of time travel, there does need to be at least a chain of counterfactual dependencies linking Tina-at-home to Tina-tanned, under standard counterfactual accounts of causation [24]. While transitivity of counterfactual dependence is not guaranteed, there is no reason to expect violations due to features such as pre-emption.

However, counterfactuals 3 and 4 cannot both be true if counterfactuals are evaluated by holding the distant present fixed. Here is why. Say in the actual world Tina shaves and is therefore beardless at the beach. Under standard semantics [25], counterfactual 3 is then straightforwardly true. However, given the requirement to hold the distant present fixed, counterfactual 4 is false. Tina-tanned must be beardless, since Tina-tanned's state is in the distant present of Tina-at home. So, if Tina-at-home were not to shave, Tina-tanned would still be beardless. Mutatis mutandi if, in the actual world, Tina does not shave: counterfactual 4 is true, but counterfactual 3 is false. Because counterfactuals 3 and 4 cannot both be true, whether Tina-tanned is bearded cannot counterfactually depend on whether Tina-at-home shaves. More generally, if the distant present is held fixed when evaluating counterfactuals, there can be no counterfactual dependence of a time traveller's distant present states on her states now.

Any counterfactual changes to the past that do occur, moreover, must always be 'put back' by the time of the present, by whatever means necessary. For example, say Tina does not shave in the actual world. If so, the following counterfactual is true:

5. If Tina-at-home were to shave, Tina-arrived-at-beach would be beardless, but her beard would grow back, reattach, or otherwise return by later that day.

While this behaviour may not be inconsistent with fundamental physical laws, it is nevertheless inconsistent with ordinary macroscopic behaviour, and it cannot be explained by the need to avoid self-defeating causal loops, nor can the implications of counterfactual 5 be contained by Lewis' [26] distinction between what an agent causes, and what would merely be true. Because of the requirements of causal continuity in time travel, Tina can cause the seemingly miraculous behaviour of her beard.

One might argue that at least one of counterfactuals 3 or 4 must be false precisely because the area outside the antecedent should be held fixed when evaluating counterfactuals.[3] If so, take Tina's case to be a way of drawing out the surprising consequences of this view—it implies violations of ordinary lawful and causal behaviour, and failures of transitivity of counterfactual dependence. The view may also imply that Tina's deliberation is out of place, since her state at the beach does not depend counterfactually on her decision now. Tina's case therefore represents a choice point—either accept the strange consequences or reject the requirement when evaluating counterfactuals in cases involving backwards time travel.

While these results are not conclusive, I will assume, for the moment, that one should not accept these strange consequences. Therefore, the requirement to hold the present fixed cannot be part of a completely general method of evaluating counterfactuals, one that delivers appropriate results for all causal structures. As I noted in Section 1, this result might not seem too surprising, since many of those who defend the requirement to hold the distant present fixed only use the requirement when evaluating counterfactuals in the actual world [1–3]. The deeper point that Tina's case demonstrates, however, is that holding the distant present fixed enforces features of counterfactual structure. The requirement rules out simultaneous counterfactual dependence *a priori* and independently of what the rest of the counterfactual and causal structure of Tina's world is like. As I will argue in the next section, this deeper point implies that holding the distant present fixed cannot be used when evaluating counterfactuals on the way to explaining causal asymmetry—contrary to certain statistical mechanical accounts.

3. Upshots for Explaining Causal Asymmetry

The first, less interesting, upshot of Tina's case is that the requirement to hold the present fixed cannot be part of a completely general method of evaluating counterfactuals. The lack of such a method will not matter for some projects. For example, altered states recipe approaches standardly hold the distant present fixed but defenders of these approaches typically aim to merely recover counterfactuals that are true of the actual world [3] (pp. 21, 31), and are sometimes explicit that their methods are not intended to cover cases involving backwards causation ([1], p. 10; [2], p. 8). Defenders can respond by simply limiting the scope of their accounts. Defenders might also respond to Tina's case by altering the requirement, such that the present is only held fixed by default, and introducing additional causal stipulations. For example, the distant present might be held fixed except for states that are in the causal future of the antecedent. See [27,28] for related proposals.[4]

The second, by far more important, upshot of Tina's case concerns how we explain causal asymmetry in the actual world. Certain statistical mechanical accounts ([6], Ch. 6; [7]; [8], pp. 234–236; [29]) use counterfactuals evaluated by holding the present fixed to explain causal asymmetry. These accounts cannot make use of the first response above. They aim to use a method of evaluating counterfactuals that, combined with the physical structure of a given world, delivers that world's causal structure. Defenders cannot assume, without explanation, that there is no backwards causation in a given case prior to determining what counterfactual method to apply. It is the counterfactual method itself

that is supposed to deliver the fact that there is no backwards causation. Because these accounts aim to explain causal structure using only non-causal features of reality, they also cannot take the second response above and include causal stipulations in the method. Moreover, as we will see, these accounts only succeed if the whole distant present is held fixed when evaluating counterfactuals. It is not enough if the distant present is held fixed by default.

The underlying concern for counterfactual-based explanations of causal asymmetry that use the requirement is that holding the distant present fixed cannot be part of a suitably temporally neutral method of evaluating counterfactuals—one that can explain causal asymmetry. To use counterfactuals to explain causal asymmetry, it is essential that the method not illicitly 'build in' features of counterfactual structure. It is a familiar point that one cannot evaluate counterfactuals by holding the past fixed when explaining causal asymmetry [24]—such a method is illicit because it rules out counterfactual dependence of the past on the present (ruling out backwards causation), independently of what the underlying physical structure of a world is like. It makes explaining causal asymmetry 'too easy'. While the requirement to hold the distant present fixed is not temporally asymmetric, it does similar harm—it enforces features of the counterfactual structure and prevents its being derived from the underlying physical structure.

Tina's case demonstrates how this occurs. Even though the causal structure of Tina's world suggests there should be interesting forms of simultaneous counterfactual dependence, relevant to her decision-making, the requirement to hold the distant present fixed rules out these dependencies *a priori* and enforces a different counterfactual structure that excludes simultaneous counterfactual dependencies—independently of the physical structure of her world. The problem raised by Tina's case is not merely that the results of the requirement are unintuitive, but that the requirement rules out simultaneous counterfactual dependence *a priori*, preventing features of counterfactual structure being explained in physical terms.

In the remainder of the paper, I will use Tina's case to argue more directly against the above statistical mechanical explanations of causal asymmetry. While, for simplicity, I will focus on Loewer's account [7], I aim to undermine what has seemed to be a promising general approach to explaining causal asymmetry.[5]

Loewer adopts a statistical mechanical method for evaluating counterfactuals. The method is used to evaluate counterfactuals where the antecedents are decisions of agents, rather than events or states more generally, and where the consequents are typically probabilities of macrostates, rather than events or states more generally. Macrostates are states of systems characterised using macroscopic language—'a mug of water at boiling point' specifies a macrostate, whereas a description of the location and velocity of all the water molecules specifies a microstate.

Here is Loewer's method for evaluating counterfactuals, put roughly. To evaluate what would be the case, were an agent to decide other than they do in the actual world, consider a partially specified counterfactual world that has the same macroscopic state as the actual world at the time of the decision, t, that started out in the same macroscopic state as the actual world, that has the same fundamental dynamical laws as the actual world, but where the microscopic state of the agent's brain at t is different—the microscopic state is such that the agent decides **D1**. Apply a statistical postulate that, roughly put, takes complete microscopically specified counterfactual worlds compatible with the above partial characterisation to be each equally probable. The probability distribution over microscopically specified counterfactual worlds implies probabilities of macrostates at any given time—which are the counterfactual consequents.

For example, to determine what would happen, were you to decide to frolic on the lawns (given that you do not in the actual world), consider a partially specified world that has you deciding, at t, to frolic on the lawns, but that has the same macrostate as the actual world at t, the same fundamental dynamical laws, and that started out in the same macrostate. If, in most of the microscopically specified counterfactual worlds consistent

with this partial specification, you do frolic on the lawns at t1, then the counterfactual 'If at t you were to decide to frolic on the lawns, the probability of your frolicking on the lawns would be high' is true.

More precisely, on Loewer's method, the conditional 'If at t I were to decide **D1** then the probability of **B** would be x' is true just in case Pr(**B**/**M**(t)&**D**(t)) = x, where **B** is a macroscopic event, **M**(t) is the macroscopic state of the world at t, **D** is the decision at t (compatible with the macrostate **M**(t)), and Pr is a chance function evaluated using statistical mechanical probabilities.[6] For example, were a particular decision **D1** to be made at t, an event **B** would occur with high probability, just in case **B**'s probability is high, given **D1**(t), and given the macrostate of the actual world at the time of **D** remains as it is in the actual world (that is, given **M**(t)).

In both the rough and precise formulation, because the probabilities used to derive counterfactuals involve conditionalising over the entire macroscopic state of the world at the time of the antecedent, that is, reasoning to a partially specified world that has the same macrostate as the actual world, the macroscopic state of the distant present is 'held fixed'.[7]

Using this method, Loewer derives a temporal asymmetry in what decision-counterfactuals are true. Assume (a) that the present contains vastly more 'macro signatures' of the past than the future. Macro signatures are local states that render other local states highly probable, given statistical mechanical probabilities [7] (p. 318)—they are often referred to in the literature as 'records' [6] (chapter 6). Records include states such as memories, recordings, fossils, etc. If we evaluate counterfactuals by holding the present macrostate fixed, any macro signatures of the actual past contained in the present also remain the same in counterfactual worlds. Because counterfactuals and macro signatures are both evaluated using the same statistical mechanical probabilities, the events they are macro signatures of in the past will also remain unchanged in counterfactual worlds. Therefore, Loewer argues, there will be no counterfactual dependence of past events recorded in the present on present decisions. Any such counterfactual dependence is ruled out by the macro signatures contained in the present.

Assume also (b) that, given our biology, small changes in our brain state can in general be probabilistically correlated with large changes elsewhere. Taking (a) and (b) together, Loewer argues there can often be significant counterfactual dependence of the future on present decisions, but that past states will not depend counterfactually on present decisions. If so, there is a temporal asymmetry in what decision counterfactuals are true—one that might explain a more general causal asymmetry ([7] (p. 297); [31]).

There are various objections one might raise to Loewer's account.[8] My point for the moment is that its success requires holding the whole distant present fixed. Here is why. Assumption (b) allows that small changes in brain states can, in general, be probabilistically correlated with macroscopic changes in the world at the same time. Say we allow that some parts of the macroscopic present outside the antecedent may be different in counterfactual worlds. Then, by (b), small changes in brain states can be counterfactually correlated with changes in the present—including changes to the macro signatures contained in the present. Because the macro signatures contained in the present will sometimes be different in counterfactual worlds, the events they record in the past will be different as well. Thus, even when there are macro signatures of the past state contained in the present, there can still be significant dependence of past events on present decisions. Moreover, given correlations between decisions and past states are macroscopic, they are the kind that we could come to know about, therefore they cannot be ruled out as unknowable ([7], p. 318; [29] p. 127). Without the requirement to hold the whole distant present fixed, Loewer's account does not begin to rule out backwards causation. The requirement to hold the distant present fixed thus plays a crucial role in Loewer's explanation of causal asymmetry.

However, as we saw with Tina's case, holding the distant present fixed enforces features of counterfactual structure, independently of the physical structure of a world. Thus, it cannot be used when deriving what the counterfactual structure of a given world is like from its underlying physical structure. One might allow Loewer to use the requirement

when deriving further features of counterfactual structure, if he provided some physical justification for the requirement—some explanation of why it accurately reflects the pre-existing counterfactual structure of a given world. At no point, however, does Loewer offer a physical justification for the requirement. There is no attempt to explain, using physical features, why there is no counterfactual dependence of the distant present at our world. Because Loewer assumes, *a priori* and without physical justification, that there is no simultaneous counterfactual dependence at our world, his purported explanation fails to adequately trace causal asymmetry back to physical features of our world. At a crucial point, Loewer's account assumes, rather than explains, what is required to derive causal asymmetry.

One might be tempted to respond that, insofar as holding the distant present fixed is part of a reasonable method of evaluating counterfactuals at our world, there can be no harm employing it when explaining causal asymmetry at our world. To be clear, I am not objecting to the method as a reasonable method of evaluating counterfactuals in our world. Indeed, the method may be reasonable in worlds in which all causal relations are aligned in the same temporal direction. However, its use in explaining causal asymmetry is a distinct issue. In an explanatory context where we are trying to derive causal structure from non-causal physical structure via counterfactuals, we cannot assume counterfactual structure that is in no way derived from physical structure.

One might instead attempt to justify the requirement to hold the present fixed and explain why it is part of a reasonable method of evaluating counterfactuals at our world. If this could be done in non-causal terms, then perhaps the requirement could still be used when giving a non-causal explanation of causal asymmetry, but the prospects do not look promising. The reason why the requirement seems to produce reasonable results in our world is because causal relations in our world are always temporally aligned. For this reason, there are no combinations of backwards and forwards causal influence that would produce simultaneous counterfactual dependence of the kind found in Tina's case. However, if we could explain why all causal relations were temporally aligned (in non-causal terms) we would have done most of the work in explaining causal asymmetry. Indeed, as we will see below, Loewer's account comes close to simply assuming there are at least some forwards causal relations. This assumption, combined with the claim that all causal relations are temporally aligned, would be enough to explain causal asymmetry without making use of the requirement to hold the present fixed or Loewer's macro signature-based explanation above.

Can Loewer justify holding the present fixed in other terms? He might be thought to do so indirectly, via an assumption about decisions [7] (p. 317):

> given the macrostate **M** of the world (including the agent's brain) the various decisions that are available to her are all equally likely. Decisions are thus indeterministic relative to the macro state of the brain and environment prior to, and at the moment of, making the decision. This indeterminacy captures the idea that which decision one makes is 'open' prior to making the decision.

While Loewer acknowledges that the assumption that 'each possible decision is equally likely is certainly false' he '[doesn't] think this simplification affects the account' ([7], p. 317, n. 39). The crucial assumption, however, is not that decisions are equally likely, but that no single decision is highly probable, given local states in the present or past. Loewer assumes, in other words, that there cannot be macro signatures of available decisions in the distant present or past. He assumes there cannot be simultaneous or previous states that reflect what decision is made in the present. If so, there cannot be cases, like Tina's, where her decision (to shave or not) is reflected in her state at the beach (beardless or not) and where holding fixed the distant present (such as whether she is beardless) leads to strange results—where changes to her beard in the past must always be 'put back' by the present. Moreover, a lack of significant correlations between available decisions and macro states in the distant present might be thought to justify holding the distant present fixed.

However, Loewer's assumption about available decisions is deeply problematic. First, as others have noted [32], the assumption is temporally asymmetric. Loewer assumes that small changes in our brain states can (in general) be probabilistically correlated with large changes elsewhere (assumption b), above). However, he also assumes that our available decisions are not probabilistically correlated with past states. If so, assuming our available decisions are probabilistically correlated with any states at other times, they will be probabilistically correlated with future states—and we have assumed that there are at least some forwards causal relations.[9] However, since Loewer assumes available decisions are not probabilistically correlated with past states, and such correlations are a requirement for counterfactual dependence, Loewer rules out backwards causal relations that are as direct as the forwards causal relations. A further problem is that if we justify the requirement to hold the present fixed by assuming all causal relations are temporally aligned (as above), the assumption that there are some forwards directed causal relations allows one to derive that all causal relations are forwards directed—independently of the counterfactual and probabilistic structure of the world.

A second problem with Loewer's assumption is it is false of decisions we make in the actual world. Provided we sometimes reliably make decisions in response to particular macroscopic events, there can be macro signatures of our decisions in the past and distant present. For example, my decision to play certain piano keys may be a macro signature of what notes I have already played [30] (p. 31), and so of what notes a sound recording in the present contains. Loewer could retreat to the position that the assumption about available decisions is merely a 'fiction' [7] (p. 317) or 'myth' [29] (p. 127). While this response deals with the second problem, it does not address the first, or a third.

The third problem is that the assumption is unreasonable, even as a fiction. The assumption implies that agents like Tina do not even fictionally count as having available decisions, merely because they take there to be states in the distant present that are macro signatures of their decisions. However, given Tina herself has no records now of her state at the beach, there is nothing in her knowledge of the past to prevent her employing the fiction. Moreover, in Tina's case, there are macro signatures in the distant present of her decisions precisely because she can control the past. Having control of the past should not rule her out as having available decisions in the fiction, and believing she has control of the past should not prevent her employing the fiction.

I suspect Loewer's assumption might have looked reasonable because we confuse direct and indirect control. It may be true that our direct control of the present is limited to a small local area, such as our brain states, but this does not imply that our indirect control of the present is similarly limited. To assume we cannot indirectly control the present is to assume something about the causal and counterfactual structure of the entire rest of the world—namely that we cannot control the distant present using the past. However, as I have argued, we are not entitled to this as an assumption when explaining causal asymmetry. While it may be true that our indirect control in the actual world is limited to the future, this cannot be simply assumed when explaining why this is so.[10]

At this point, one might be tempted to give up on the project of explaining causal asymmetry. If one does not use counterfactuals to derive causal structure from non-causal structure, one can adopt a causal method of evaluating counterfactuals ([3,13,27,28,34]). Causal approaches could be used to derive counterfactuals in the actual world. They could also be used to deliver the intuitive results in time travel cases like Tina's, provided they didn't hold events in the causal future of the antecedent 'fixed'.[11] However, adopting only a causal approach to evaluating counterfactuals would mean giving up on the project of explaining causal asymmetry in counterfactual terms. For reasons explored [4–6] and elsewhere, this would be to abandon an otherwise promising approach to explaining causal asymmetry in scientific terms and unifying a range of temporally asymmetric phenomena.[12]

Moreover, there are alternatives. To explain an asymmetry of decision counterfactuals in the actual world, what we need to do is rule out counterfactual dependence in cases where an agent's decision is (or is taken to be) evidence of, or probabilistically correlated

with, the states she is responding to—cases like the piano player above. We do not need to rule out counterfactual dependencies in other kinds of cases. Ruling out counterfactual dependence concerning an agent's responses is the approach taken by [9,35–37]. According to these broadly physicalist agent-based approaches, the evidence the agent has while deliberating prevents their decision being a means to raise the probability of states they are responding to. For this reason, the piano player's past playing does not depend counterfactually on their decision now—or at least not in a way that would amount to the agent's decision influencing, controlling, or causing their past playing.

These response-focused accounts rule out backwards counterfactual dependence in the piano player case and provide an alternative route to explaining causal asymmetry in the actual world, but they do not vindicate a general requirement to hold the distant present fixed when evaluating counterfactuals. They allow for simultaneous counterfactual dependence in Tina's case. Precisely because they do not hold the distant present fixed, they do better when applied to worlds with complex causal structures. These methods work because they are sensitive to local probabilistic and counterfactual dependencies, and do not employ global requirements such as holding the present fixed. While these approaches still face difficult choices concerning precisely what to hold fixed, particularly in cases involving causal loops (see [18] for discussion), the local nature of these approaches makes them better candidates for explaining, in physical terms, why our world has the causal structure that it does.

There are no doubt other alternatives to explore. Regardless, the lesson of Tina's case is that if we are to explain even a global causal asymmetry in our world in physical terms, the method used cannot employ the global requirements to 'hold the distant present fixed'. Such a requirement presumes features of counterfactual structure and prevents their being explained in physical terms. A promising alternative is to use methods that are more sensitive to local structure, including those that only rule out dependencies in case where an agent's decisions are responses to events.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 The case will not rely on gender-swapping or anything like that.

2 Plausibly, her having no knowledge of her state at the beach is a requirement on her reasonably deliberating about her decision [19]. One might argue that causal loops involving the agent's decisions are unavoidable in cases of backwards time travel, even if Tina travels into the far distant present. If Tina's case involves a causal loop, it is perhaps less surprising that methods of evaluating counterfactuals fail. See [11,13,20–23] for discussion for some of the difficulties evaluating counterfactuals in cases involving causal loops. My concerns with holding the distant present fixed are unrelated to causal loops.

3 Ideally, one would also want some independent motivation for the requirement. The standard Lewisian motivation [24] is to recover our intuitive judgements, but that is no help when the intuitions are in question or favour an alternative.

4 While [3] (pp. 30–31) might be thought to attempt a non-causal solution, his solution is temporally asymmetric and has a causal flavour, particularly regarding his talk of 'infection'.

5 Loewer [29] uses probabilities in place of counterfactuals, but uses the same requirements for how each are evaluated and takes counterfactual structure to derive from probabilistic structure [29] (p. 132). Albert's original account [6] (chapter 6) is ambiguous but is often interpreted as holding the distant present fixed [30] (p. 27). Albert confirms (private communication) that this is what he had in mind. Kutach [8] (pp. 234–236) uses the requirement to explain why we cannot influence the past by means of our forwards influence. While Kutach accepts that this asymmetry of influence may not hold in time travel scenarios [8] (p. 229), he does not take the requirement itself to be problematic when explaining temporal asymmetries.

6 Statistical mechanical probabilities are derived from taking the Lebesgue probability measure over microstates compatible with the low-entropy macrostate of the early universe—the 'Past Hypothesis'—and conditionalising over later macrostates [7] (p. 317).

7 Whether the macrostate [7] or the microstate [29] outside the antecedent is held fixed will not matter to my arguments. If holding the macrostate fixed is problematic, as Tina's case suggest, then holding the microstate fixed is also problematic.

[8] What if there are no macro signatures of a past (or future) event contained in the present? Loewer [7] (p. 318) responds that, in that case, the past (or future) event will not probabilistically depend on the present decision. Loewer's explanation of the asymmetry fails, however, if the decision is the only record of the past event in the present [30]. I discuss this kind of case below.

[9] At least on a standard Lewisian counterfactual account of causation [24]. Again, Loewer is not explicit about the precise relation between counterfactuals and causal relations.

[10] Could Tina's case be ruled out because it implies violations of thermodynamic asymmetries? It is controversial whether time travel (along time-like curves) implies such violations [33] (p. 137), but, to make this response sufficiently general, one would need to argue that backwards causation implies violations of thermodynamic asymmetries, and it is precisely to be shown, not assumed, that the direction of causation is to be explained in statistical mechanical or thermodynamic terms when giving statistical mechanical explanations of causal asymmetry.

[11] Causal methods of evaluating counterfactuals face difficulties dealing with causal structures such as causal loops—see [11,13,20–23] for discussion. Note that the standard ways of dealing with counterfactuals in cases of causal loops will not help Loewer either—standard accounts presume either temporal asymmetry [11,21–23] or are causal [13,20]. See [18] for discussion.

[12] A similar point holds for accounts that use probabilities rather than counterfactuals to explain causal asymmetry [29]. Causal methods of evaluating probabilities cannot be used if the project is to explain causal asymmetry using probabilities.

References

1. Collins, J.; Hall, N.; Paul, L.A. Counterfactuals and Causation: History, Problems, and Prospects. In *Causation and Counterfactuals*; MIT Press: Cambridge, MA, USA, 2004; pp. 1–58.
2. Paul, L.A.; Hall, N. *Causation: A User's Guide*; Oxford University Press: Oxford, UK, 2013.
3. Maudlin, T. *The Metaphysics within Physics*; Oxford University Press: Oxford, UK, 2007.
4. Field, H. Causation in a Physical World. In *The Oxford Handbook of Metaphysics*; Loux, M.J., Zimmerman, D.W., Eds.; Oxford University Press: Oxford, UK, 2003; pp. 435–460.
5. Fernandes, A. The Temporal Asymmetry of Causation. *MS.* **2022**.
6. Albert, D. *Time and Chance*; Harvard University Press: Cambridge, MA, USA, 2000.
7. Loewer, B. Counterfactuals and the Second Law. In *Causation, Physics, and the Constitution of Reality*; Price, H., Corry, R., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 293–326.
8. Kutach, D. *Causation and Its Basis in Fundamental Physics*; Oxford University Press: Oxford, UK, 2013.
9. Price, H.; Weslake, B. The Time-Asymmetry of Causation. In *The Oxford Handbook of Causation*; Beebee, H., Menzies, P., Hitchcock, C., Eds.; Oxford University Press: Oxford, UK, 2009.
10. Kutach, D. The Physical Foundations of Causation. In *Causation, Physics, and the Constitution of Reality*; Price, H., Corry, R., Eds.; Oxford University Press: Oxford, UK, 2007; pp. 327–350.
11. Lewis, D. The Paradoxes of Time Travel. *Am. Philos. Q.* **1976**, *13*, 145–152.
12. Arntzenius, F.; Maudlin, T. Time Travel and Modern Physics. In *The Stanford Encyclopedia of Philosophy (Winter 2013 Edition)*; Zalta, E.N., Ed.; Metaphysics Research Lab., Stanford University: Stanford, CA, USA, 2013; Available online: https://plato.stanford.edu/archives/win2013/entries/time-travel-phys/ (accessed on 18 February 2022).
13. Wasserman, R. *Paradoxes of Time Travel*; Oxford University Press: Oxford, UK, 2018.
14. Price, H. *Time's Arrow & Archimedes' Point*; Oxford University Press: New York, NY, USA, 1996.
15. Tooley, M. Backward Causation and the Stalnaker-Lewis Approach to Counterfactuals. *Analysis* **2002**, *62*, 191–197. [CrossRef]
16. Wasserman, R. Lewis on Backwards Causation. *Thought* **2015**, *4*, 141–150. [CrossRef]
17. Cusbert, J. Backwards Causation and the Chancy Past. *Mind* **2018**, *127*, 1–33. [CrossRef]
18. Fernandes, A. Time Travel and Counterfactual Asymmetry. *Synthese* **2021**, *198*, 1983–2001. [CrossRef]
19. Fernandes, A. Freedom, Self-Prediction, and the Possibility of Time Travel. *Philos. Stud.* **2020**, *177*, 89–108. [CrossRef]
20. Vihvelin, K. What time travelers cannot do. *Philos. Stud.* **1996**, *81*, 315–330. [CrossRef]
21. Smith, N.J.J. Bananas enough for time travel? *Br. J. Philos. Sci.* **1997**, *48*, 363–389. [CrossRef]
22. Sider, T. Time travel, coincidences and counterfactuals. *Philos. Stud.* **2002**, *110*, 115–138. [CrossRef]
23. Ismael, J. Closed Causal Loops and the Bilking Argument. *Synthese* **2003**, *136*, 305–320. [CrossRef]
24. Lewis, D. Counterfactual Dependence and Time's Arrow. *Noûs* **1979**, *13*, 455–476. [CrossRef]
25. Lewis, D. Causation. *J. Philos.* **1973**, *70*, 556–567. [CrossRef]
26. Lewis, D. Are We Free to Break the Laws? *Theoria* **1981**, *47*, 113–121. [CrossRef]
27. Edgington, D. Counterfactuals and the Benefit of Hindsight. In *Cause and Chance: Causation in an Indeterministic World*; Dowe, P., Noordhof, P., Eds.; Routledge: New York, NY, USA, 2004; pp. 12–27.
28. Schaffer, J. Counterfactuals, Causal Independence, and Conceptual Circularity. *Analysis* **2004**, *64*, 299–309. [CrossRef]
29. Loewer, B. Two accounts of laws and time. *Philos. Stud.* **2012**, *160*, 115–137. [CrossRef]
30. Frisch, M. Does a Low-Entropy Constraint Prevent Us from Influencing the Past? In *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*; Ernst, G., Hüttemann, A., Eds.; Cambridge University Press: Cambridge, UK, 2010; pp. 13–33.
31. Loewer, B. The Mentaculus Vision. In *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature*; Allori, V., Ed.; World Scientific: Singapore, 2020; pp. 3–29.

32. Beebee, H. Causation, projection, inference and agency. In *Passions and Projections: Themes from the Philosophy of Simon Blackburn*; Johnson, R.N., Smith, M., Eds.; Oxford University Press: New York, NY, USA, 2015; pp. 25–48.
33. Callender, C. *What Makes Time Special*; Oxford University Press: Oxford, UK, 2017.
34. Bennett, J. Counterfactuals and Temporal Direction. *Philos. Rev.* **1984**, *93*, 57–91. [CrossRef]
35. Blanchard, T. Causation in a Physical World. Ph.D. Thesis, Rutgers University, New Brunswick, NJ, USA, 2014.
36. Albert, D. *After Physics*; Harvard University Press: Cambridge, MA, USA, 2015.
37. Fernandes, A. A Deliberative Approach to Causation. *Philos. Phenomenol. Res.* **2017**, *95*, 686–708. [CrossRef]

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Philosophies Editorial Office
E-mail: philosophies@mdpi.com
www.mdpi.com/journal/philosophies

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34

www.mdpi.com

ISBN 978-3-0365-5540-9