Probability and Coincidence in Time Travel Scenarios

Abstract: In this essay, I argue that certain classic time travel scenarios (e.g. a person travelling back into their recent past) are very improbable. This is, I argue, because the greater the disagreement between that which a scenario's records imply happen during some period, and that which its dispositions imply happen during that same period, the more improbable the scenario is. I argue that the classic backwards time travel scenarios almost invariably involve systems and spacetime structures such that this disagreement is serious. In the course of defending these conclusions, I provide a precise definition of time travel scenarios, and engage in detailed discussion of the role objective chance and spacetime structure plays in determining the probability of the events (and specifically 'coincidences') required by time travel scenarios.

Contents

1. Introduction

Thanks to science fiction, most people are today familiar with the paradoxes of time travel: a person goes back in time and does something which prevents their time travelling in the first place (or even existing at all). If one assumes no parallel timelines and no special paradox-preventing forces, it quickly becomes apparent that these paradoxes represent a prima facie serious problem for time travel. For if it *is* possible to travel backwards in time, it is very hard to see what would possibly prevent such paradoxes from occurring, leading some to claim that time travel must just be impossible.

As we shall see, this particular conclusion is unwarranted; time travel may be possible even though paradoxes are not. Nonetheless, the manner in which time travel worlds would have to consistently avoid paradoxes without any centralised mechanism for doing so seems to require the presence of systematic uncaused correlations, and this has prompted various arguments to the effect that even if time travel is possible, it probably does not occur. At the core of this essay is an argument of this exact kind. However, my approach differs from previous arguments against time travel in its emphasis on there being different *types* of time travel scenario, and that certain features of a scenario affect the extent to which it requires these uncaused correlations. The ultimate aim here is thus not to make any ruling for or against time travel in general, but to instead present a precise method by which the (upper bounds of) probabilities of individual time travel scenarios may be ascertained.

Before reaching that point, though, we need to know what is actually meant by *time travel*. Note that throughout this essay by *time travel* I mean only *backwards* time travel: travel into the *past*. Time travel into the future, while interesting, does not fundamentally differ from the mundane temporal persistence we all achieve every day, and so I shall leave it completely out of the discussion.

In *§2: What is Time Travel?* I present an original definition of time travel which captures the essence of the notion in terms of trajectories. I then introduce the idea of a *scenario* as the basic unit of description throughout this essay. I show the relation between my definition of a time travel scenario (TTS) and the existence of closed timelike curves (CTCs) in spacetime. It is at this point which my account makes contact with the mainstream discussion of time travel, which acknowledges both that CTCs are necessary for time travel, and that they are nomologically possible. Finally, I address the fact

that my definition of TTSs is unintuitively permissive. I simply accept this feature of the definition, since the aim is not to establish whether *any* TTSs occur, but rather what kind do. I end §2 with a discussion of what extra constraints physics may place on what TTSs occur, arguing that these constraints are not sufficiently well understood to be currently of much use.

With the basic framework established, I will move on to the argument itself. *§3: The Grandfather Arguments* consists of discussions of two of the literature's most prominent arguments that TTSs do not occur. Indeed, the first, the *modal grandfather argument* (MGA), is actually an argument that TTSs are outright impossible. After presenting the MGA, I demonstrate that it is deflated once a more nuanced interpretation of possibility is adopted. However, while this line of defence is effective at undermining the MGA, it highlights that TTSs tend to entail the occurrence of apparently improbable uncaused correlations (UCs). These UCs are the focus of the second argument, Horwich's *probabilistic grandfather argument* (PGA-1), which purports to show that TTSs, due to their entailing these UCs, are generally very improbable. In the form Horwich presents it, though, the argument is rather vague and vulnerable to a variety of criticisms, one of which, attributable to Smith, I present. I finish by arguing that it is unnecessary to frame the argument in terms of UCs and Horwich's doing so undermines the force of PGA-1.

In *§4: Records, Dispositions and Self-Subversion*, I introduce records and dispositions: the means by which a system carries information about its past and its future respectively. Understanding records and dispositions is important for understanding what makes TTSs unique over other kinds of scenario. I then present PGA-2 as establishing the improbability of certain TTSs and introduce the idea of self-subversion as a simple way of understanding this improbability. Finally, I consider how a certain criticism of PGA-1 carries over to PGA-2.

§5: Chance sees the introduction of chance alongside probability. The idea is that parts of PGA-2 are best justified by appeal to chance, construed broadly as a physical quantity (unlike probability, which I construe ultimately as a rational constraint on degrees of belief). Spelling this out requires a precise theory of chance and how it relates to probability. I present and discuss two such theories: Lewis' temporal theory, and my own outcome-setup theory, arguing that the latter is more generally applicable and better suited for the task at hand. I then discuss the best system interpretation of chance,

which is adopted for the remainder of §5. Finally, I present the *chance argument* (CA), which establishes the contentious premise of PGA-2. The rest of the section is spent justifying CA in detail.

Sections §3-5 revolve round a narrow and contrived set of examples. In *§6: Self-Subverting Scenarios* moves are made to generalise the reasoning of the earlier sections to a more general and typical set of TTSs. I identify the main determining factor of whether a scenario self-subverts to be what I call its *evidential overlap*, after which I discuss what features of a scenario mainly determine this overlap. Finally, I examine some typical systems one finds in the universe and use the notion of overlap to assess how probable it is that such systems might time travel.

In *§7*, I make some concluding remarks and discuss where one might take the discussion next, and in *§8: Appendix* I prove the mathematical results I rely upon in PGA-1 and PGA-2.

So, in sum, the purpose of this essay is to present an old argument in an new, more precise form. Doing so minimises the fantastic feel common in discussions of time travel, allowing one to see more easily that the justification for the argument is very simple and based on fairly obvious and mundane premises. The form the argument takes also highlights that *time travel* is no more homogenous as a concept than *travel*, and one must address the details of individual scenarios, rather than aiming straight away to establish sweeping conclusions about time travel in general.

2. What is Time Travel?

In this section, I set up what distinguishes time travel scenarios from other types of scenario, and discuss how they square with current relativistic physics.

2.1 Time Travel Trajectories

This subsection will consist of a lot of definitions, but they will all be relied upon throughout the essay, so it is worth the rigmarole. The basic idea of a time travel trajectory is that it is a trajectory which, were a system to travel along it, the system would end up in its own past.

But first, all subsequent discussion takes place against a spacetime geometric backdrop, so a quick review of the core concepts will be necessary. (Throughout this essay, the general theory of relativity will be assumed, as it is currently our best theory of space and time.) A spacetime is an ordered pair $\langle M, g_{ab} \rangle$ where M is a connected four-dimensional differentiable manifold, and g_{ab} is a Lorentz-signature metric defined everywhere on M (and which satisfies Einstein's field equations for some stress-energy tensor). All spacetimes dealt with here are time-orientable, meaning that for every point $p \in M$, the null cones for that point have a *past* and *future* lobe. A *trajectory* is just a timelike curve (i.e. a curve whose tangent vector is everywhere timelike – falling within the future null lobe).[1]

Take a system to be a massive (photons cause unnecessary complications) relatively localised object, such that we can suppose that it has a unique *centre*. A system s follows a trajectory C iff every point $p \in C$ is the centre point of the system at some instant. Note that while there is presumably some maximal trajectory followed by a system (i.e. the trajectory over its entire lifetime), we will generally be discussing a system's following only some small section of this trajectory. Since we are assuming that spacetime is time-orientable, trajectories will always have a *forwards* time direction and a *backwards* time direction, within the future and past lobes respectively of each $p \in C$. All the usual

[1] See Malament, 2012, ch.1-2 for more details.

dynamical changes occur in a system as it progresses forward along the curve: entropy increases, nuclei decay, chemical reactions occur, etc.[2]

For any two points p and q on a trajectory, then, p is *earlier* than q iff q is forward of p on the trajectory. p is *later* than q iff q is backwards of p on the trajectory. The *earliest* and *latest* points on a trajectory are the two points such that no point on the trajectory is earlier or later, respectively. Clearly, if the trajectory is closed (i.e. comes back to meet itself), there will be no earliest or latest point. To keep things simple, I will assume that no closed trajectories are followed by any system.

Finally, p is in the *past* of q iff $p \in I^-(q)$, where $I^-(q)$ is the set of all points x such that there is some timelike curve from x to q. In general, I shall also say of two regions Γ and Σ that Γ is in the *past* of Σ iff for some point $\gamma \in \Gamma$ and $\sigma \in \Sigma$, $\gamma \in I^-(\sigma)$, thus using *is in the past of* ambiguously between relating points and regions.

Now to bring in time travel. A trajectory C is a *time travel trajectory* (TTT) iff the *latest* point on C is in the *past* of the *earliest* point on C. (Note that this means trajectories like C_2 (*see figure 1*) are not TTTs, despite travelling through the past. Still, all such trajectories *contain* a TTT (e.g. C_1 ($\subseteq C_2$)), so are essentially captured by the definition.)

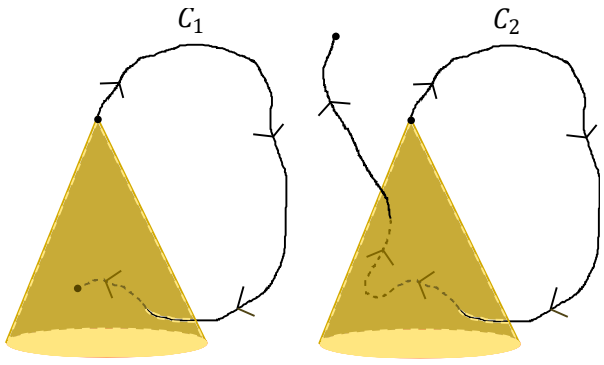

Figure 1 (colour not required): Two overlapping spacetime trajectories.

[2] It is somewhat of a puzzle within the philosophy of physics *why* this characteristic change in entropy etc. invariably occurs always in the same (forward) temporal direction. I will not address this question here, and will just assume that systems *do* in general show this characteristic evolution as they move forward along trajectories.

2.2 Time Travel Scenarios

TTTs capture the basic essence of time travel. However, saying much that is interesting about time travel at a broader level requires a compact way of specifying both a trajectory and the properties of the system which follows it.

To this end, I introduce scenarios.

Scenario: A scenario K is a proposition that trajectory C_K (with earliest point p_K^0 and latest point p_K^1) is followed by a system k, and that k has properties in the set $\aleph_K$ at p_K^1.

$\aleph_K$ may consist of only a few vague properties (e.g. is about 10m long; is made of metal), or it may be a highly extensive and exact specification of the system's state, or anything in between. Basically, a scenario says that a system with certain features is in a certain position, and has got there by a certain route through spacetime.

Time Travel Scenario (TTS): A scenario K is a *time travel scenario* iff C_K is a TTT.

Scenarios provide a neat way to formulate the aims of this essay. I will be attempting to establish which types of TTS, if any, we are justified in thinking occur. (Note that I will often talk about scenarios as if they were events. This is purely for stylistic reasons; scenarios are propositions, and my saying that one occurred/exists is just a way of saying that it is true.)

2.3 Chronology Violating Spacetimes

What does physics have to say about all this? Answering this requires noticing that any TTS will entail the existence of closed timelike curves (CTC) in spacetime, i.e. timelike curves which return to their starting point. This is because a TTS, by entailing the existence of some TTT, entails that there are two points p_K^0 and p_K^1 (the trajectory's endpoints) such that there is both a timelike curve from p_K^0 to p_K^1 (i.e. C_K) and one from p_K^1 to p_K^0 (because $p_K^1 \in I^-(p_K^0)$). Thus, it is a necessary condition for the truth of any

TTS that spacetime does indeed contain some CTC. A spacetime containing CTCs is called a *chronology violating spacetime* (CVS)[3].

There are several kinds of CVSs which are consistent with Einstein's field equations. The first such CVS was discovered by Kurt Gödel in 1949 and is a spacetime of a universe with a global homogeneous rotating mass density (and nonzero cosmological constant)[4]. Every point in the Gödel spacetime has a CTC passing through it, meaning TTSs could in principle be very numerous in such spacetimes. It is worth noting, though, that there is some good reason to think our universe is *not* Gödelian.

A simpler and less implausible CVS is produced by simply taking Minkowski spacetime and identifying two global time slices. Call this a *chronocylindrical spacetime* (my terminology). This will also have CTCs through every point.

Less contrived is the Kerr spacetime, which is the spacetime surrounding a rotating, uncharged axially-symmetric black hole. Only specific regions in the proximity of the black hole will contain CTCs. This is probably the most realistic CVS, because rotating black holes probably exist in our universe.

Finally, there are spacetimes which contain wormholes. If one 'end' of a wormhole (there is really no strict point at which a wormhole ends and the rest of space begins) is in the past of the other, then there will be CTCs which pass through the wormhole.

Wormholes *are* solutions of the field equations, but their existence is still highly speculative. Even so, what the existence of these spacetimes shows is that TTSs cannot be ruled out on purely nomological grounds. So far as we know, the spacetime of our universe may well be chronology violating in any of the above ways. Still, more is required for an interesting TTS than just the presence of CTCs. Showing this first requires having some idea of what makes a TTS interesting in the first place.

[3] See Smeenk and Wüthrich, 2011, §3 for extensive discussion of CVSs.
[4] See Malament, ch.3

2.4 Interesting Time Travel

It probably will not have escaped readers that on my definition of them, TTTs are just any section of a CTC. That is, so long as the trajectory in a scenario is part of *some* CTC, then the scenario is a TTS. This in effect makes my definition of TTSs extremely permissive, for there are many cases in which seemingly ordinary scenarios are TTSs. For instance, if our spacetime is indeed a Gödel spacetime, then my trip to the kitchen for a cup of tea will count as my time travelling. Or in more precise terms, that I go from my room to my kitchen is a TTS. This is because every trajectory in a Gödel spacetime, including my trip, is a section of a CTC. Similarly for the chronocylindrical spacetime. In both spacetimes, *all* scenarios are TTSs.

Wormhole spacetimes do not in general have CTCs through every point. Still, so long as my trip takes place entirely within the future and past light cones of the two ends of a wormhole, even if those ends are billions of years in my past and future, then I will be time travelling.

This observation might incline one to think that my definition is somehow inadequate; that I have missed some necessary condition. However, since the room-kitchen trajectory could be (in a Gödel spacetime) continuously deformed into a trajectory which takes me back to eleventh century Hastings, it is very difficult to see how some principled distinction just in terms of spacetime geometry could be drawn between the two scenarios, despite their clear differences. Considerations like these suggest that, rather than trying to shoehorn our vague expectations of what time travel should look like into a definition of TTSs, we should treat the permissive definition as a precise tool for establishing general principles for dealing with TTSs of *all* kinds. One can then apply these principles to scenarios they consider clear cases of time travel while ignoring all those they consider artefacts of my definition. Thus, the ultimate goal of the essay should not be seen as establishing whether *any* TTSs occur (for they may be trivially ubiquitous, depending on the spacetime), but rather on *what kinds* of TTSs occur.

2.5 Setting the Physics Aside

Even if our spacetime is a CVS, it is clear that what sorts of TTSs can occur will be constrained by the specifics of the CVS. For instance, consider a just the minimal TTS in which a system follows a trajectory which brings it back to reasonably close to its starting time and position.

In a Gödel spacetime, this will require the system to undergo a large constant acceleration, and it is uncertain that any rocket-like device could carry enough fuel to complete the journey[5]. Similarly for the chronocylindrical spacetime: the system would have to persist long enough for its trajectory to come back to near where it started, and if the cycle time of the spacetime is very long (and it is plausibly at least a few billion years, given what we have observed), the system will be unlikely to survive long enough to complete its journey. For a Kerr spacetime, the system will have to get into the vicinity of the black hole without being destroyed or falling in. In the case of wormholes, a wormhole will have to be of a large enough size and last long enough to allow the system to pass through intact. Plus, the two ends of the wormhole have to be sufficiently close together for the system to reach its starting region.

The issue with all of this is that we really have very little information about any of these parameters at this stage. We do not know whether any propulsion system is possible which could successfully navigate the Gödel spacetime; or if a system could survive billions of years; or if a wormhole of sufficient size and lifetime could be found or created, and whether the ends would be situated in an interesting way.

Thus, the strategy adopted in this essay is to completely set all of this aside and to simply assume that TSSs are completely unconstrained by the physics of time travel itself. This is not to say that I will be ignoring physics; the idea is to explore how more familiar physical considerations which have nothing directly to do with time travel constrain the occurrence of TTSs. This is thus a *further* constraint, on top of whatever constraints the time travel physics actually does place. In the next section we will see how some more familiar, non-cosmological facts might constrain time travel scenarios.

[5] See Malament, 1985

3. The Grandfather Arguments

In this section, I discuss and criticise two influential arguments: one that TTSs are *impossible*, and one that TTSs are *improbable*, arguing that while the latter is formulated inappropriately, it is basically correct.

3.1. The Kevin Scenario

There have been a variety of arguments to the conclusion that certain classes of TTS do not occur, and which are completely neutral on the physics required for time travel. This section will be focussed on one argument in particular: that presented by Paul Horwich[6], which claims that TTSs do not in general happen because they rely on the occurrence of certain kinds of improbable events. There are several considerations which motivate Horwich's claim, and these are best brought out by first considering an older argument, which claims that certain TTSs are not just improbable, but impossible.

It should be noted that neither this argument nor Horwich's are usually presented in terms of particular TTSs, but rather as aiming to establish results about the occurrence of time travel in general. Despite this, they tend to proceed by means of a rather narrow (and contrived) set of example TTSs, showing these to be problematic in certain ways, and then making some broad extrapolation to the general case. This should strike one as not a particularly reliable strategy, and it is not. I will thus here treat these arguments only insofar as they rule out the specific sort of examples they adduce. The reader is free to consider what more general conclusions these arguments may support, but I will mostly just be spelling out their salient rationale as applied to a few examples. Only later (§6) will I attempt to extrapolate this rationale to the general case.

Much of the time travel literature concerns itself with the classic sci-fi scenario of a person travelling a few decades into the past. In-keeping with this tradition, I introduce the Kevin scenario K. Kevin, a normal 21st century man, travels from 2019 Oxford to 1929 Portsmouth (say, by jumping through a conveniently placed wormhole) with the intention of killing his (infant) grandfather, Lloyd. Note that the precise trajectory C_K and endpoints are left mostly implicit. The specified set of properties

[6] Horwich, 1987, p.116-23

$\aleph_K$ consists of just those properties a normal 21st century man would have (e.g. certain clothes, a smartphone, a human genome, etc.) plus the intention of wanting to kill Lloyd (and accompanying preparatory information). Now let's see exactly why this is supposed to be problematic.

3.2. The Modal Grandfather Argument

The modal grandfather argument (MGA) constitutes probably the most well known and most natural objection to time travel. Applied to the Kevin scenario, the MGA runs as follows. Assume K. Thus, Kevin arrives in 1929 Portsmouth with precise information about baby Lloyd's location, the desire to kill the child, and the will and skill to do so. Thus, by any reasonable standard, Kevin is able to kill Lloyd. But, of course, Kevin cannot kill Lloyd, for Lloyd's survival is guaranteed by Kevin's existence in the first place (assuming no resurrections, etc.). Nobody is both able and unable to do something, so K is false; Kevin never embarks on the homicidal journey.

(One might immediately object that the scenario is pretty unrealistic anyway. Why would anyone want to kill their grandfather? I will address this sort of worry in §3.5 when discussing Horwich's argument.)

The argument is certainly intuitive insofar as it seems that once Kevin has successfully made it to 1929, his local circumstances would be no different from that of an identical person who had not time travelled, and who had all the same information and intent. That person would certainly succeed, and thus it seems so would Kevin. The only way to resolve the incongruity is to deny that Kevin would ever get to 1929 in the first place. And since ability to time travel ought not to depend on the intentions of the traveller, the MGA seems to undermine all TTSs where someone travels back into their past.

The argument falls down, however, once one adopts a more precise notion of possibility (and hence ability), as pointed out by Lewis[7]. Lewis claims that what is possible, given a particular scenario, will depend on how much of that scenario is held fixed when considering alternate possible worlds. Something is possible if there exists a possible world both in which that thing happens and in which the fixed facts (true propositions) are true. When we conclude that Kevin *can* kill Lloyd, we are holding

[7] Lewis, 1976, p.150

fixed the fact that Kevin journeys from 2019 to 1929 and has the information, inclination and will sufficient to kill Lloyd. There is then most certainly a possible world where these facts are true, and in which Kevin kills Lloyd. But in these worlds Lloyd is not Kevin's grandfather – presumably Kevin makes a mistake and targets the wrong person. (It is possible that we do not actually hold fixed even that Kevin is a time traveller, in which case there are worlds in which Kevin kills Lloyd but has actually hallucinated the entire time travel story and was really born in 1898.)

It is only when we hold fixed all this *plus* the fact that Lloyd is Kevin's grandfather that it becomes impossible for Kevin to succeed in the murder. But this need not imply anything about time travel. When considering what is possible, we can almost always make something which does not actually happen come out as impossible by holding fixed enough of the surrounding facts. For instance, it is possible that I have pasta tonight, even though I will not. I have the ingredients, the culinary ability, a working stove, etc. But if we hold fixed enough facts about my brain state and environment, we can close all possible ways which I might have come to have pasta, and thus it will be impossible that I do so.

The MGA requires holding fixed different facts at different points in the argument, therefor equivocating on what it is possible for Kevin to do. If one holds fixed that Lloyd is Kevin's grandfather consistently throughout the argument, one never infers from K that it is possible that Kevin succeeds, and thus no contradiction arises.

3.3. Commonplace Reasons

Lewis' analysis makes plain that the MGA fails to show that K entails a contradiction, and so fails to show that K is false. Nonetheless, Lewis' analysis also demonstrates more clearly than ever that, given that Lloyd *is* Kevin's grandfather, it is impossible that Kevin succeeds. This means that in every world in which K occurs, something prevents Kevin from killing Lloyd.

What is this force which stands poised to intervene whenever Kevin or someone like him attempts to do the impossible? God? The time lords? Or some yet more mysterious force? Any such governing

entity would be a strange and implausible thing indeed, and if K entails the existence of such an entity, then this provides good reason for doubting that K occurs.

As Lewis points out, though, no such entity is required[8]. It is unnecessary to postulate some general consistency saving force in order to account for the impossibility of Kevin's success. Instead, in each world in which K is true, we need only suppose that some minimally improbable event happens to foil Kevin's plan. What the minimally improbable foiling event is will depend upon details of the world and scenario. If Kevin arrives at a time of social unrest, it might be that a riot blocks his route. If he is unskilled with a rifle, he may miss the shot. The point is that Kevin's failure can be explained by appeal to the exact same sort of *commonplace reasons* by which one's failure to do *any* particular thing might be explained.

3.4. Horwich: The Probabilistic Grandfather Argument

So, we should assume that what foils Kevin in a given scenario is the most probable commonplace event which would result in his failure in the scenario. Still, it may be that the most probable reason is still rather improbable. This is especially the case if Kevin repeatedly tries, unperturbed by each failure, to kill Lloyd. Even if the foiling event in each case is not altogether improbable, the occurrence of the entire sequence of events *is* very improbable (the more so the greater the length of the sequence). Horwich's claim is that since K entails that such an improbable string of foiling events does occur, K is at least as improbable as the string is. I will spell this out more precisely shortly.

First, though, it must be made clear what is being claimed when I say that something is probable or improbable. Throughout this essay, I interpret probability as a prescription of rational degree of belief. It is not subjective in the sense that it varies from person to person, but it is fundamentally about subjective (personal) probabilities. Thus, $P(A|B) = x$ means: *if one were to (fully) believe that B, one ought to believe to degree x that A*. This is in line with my framing the goal of this essay as being to establish what TTSs we are justified in thinking occur. Probability, more or less, measures the degree of this justification.

[8] Lewis, 1976, p.149-50

To some, this might seem like changing the subject; that we are now talking not about time travel but instead our own beliefs. Others might doubt that such a normative interpretation of probability makes sense in the first place. At this stage I would only ask that objectors of both kind bear with me. Initially, I will be using probability construed in this manner simply as a way of making precise comparisons between different types of evidence for various propositions, which may give the use of probability an air of emptiness (just attaching numbers for the sake of it). All of §5, however, will be concerned with connecting this notion of probability to more concrete notions (e.g. relative frequencies), whereupon the role of probability becomes much clearer. As for the latter concern, I leave any problems with normativity unaddressed, since they have no special bearing on the topic of time travel. Note also that I do not claim that any authors I discuss here share this interpretation of probability, but I do think their claims can generally be best understood in terms of it.

With that out of the way, we can reconstruct Horwich's argument. (Note that Horwich presents it as an argument against *time travel* as a general phenomenon, rather than a particular TTS. The argument is much clearer, though, if one applies it to a particular TTS and *then* attempts to generalise it.) Where Q: *a string of foiling events occurs*, and E is all the various background information we have (containing laws, generalisations and particular historical facts), my reconstruction goes:

The Probabilistic Grandfather Argument 1 (PGA-1):

P1: $P(Q|KE) = 1$

P2: $P(Q|E) \ll 1$

P3: $P(K|E) \leq P(Q|E)/P(Q|KE)$ (See Appendix §8.1)

C1: $P(K|E) \ll 1$

The argument is clearly valid. P1 is justified by the points considered in §3.3; K's truth entails (or at least strongly implies) Kevin fails, which itself requires (in the absence of some policing force) the occurrence of the foiling events. P3 just follows from the definition of conditional probability. P2 is a little more tricky to justify, and indeed, much of this essay will be spent discussing this claim.

Horwich justifies P2 principally by appeal to the *Principle of V-Correlation*:

> if events of type A and B are associated with one another, then either there is always a chain of events between them…or else we find an earlier event of type C that links up with A and B by two such chains of events. What we do not see is…an inverse fork—in which A and B are connected only with a characteristic subsequent event, but no preceding one. (Horwich, 1987, 97–8)

I take this to mean basically that uncaused correlations (UCs) are improbable. And, Horwich argues, the foiling events qualify as UCs; whenever Kevin sets out to kill Lloyd, some foiling event occurs: he slips on a banana peel, he forgets his wallet, his gun jams, etc. These events are precisely correlated with his murder attempts. Yet, there is no causal basis for this correlation. This idea can be made clearer by considering the case in which there *is* some consistency-policing force like the Time Lords. Regardless of how the Time Lords are supposed to work, that their arrival to foil Kevin is correlated precisely with his murder attempts is perfectly explicable in causal terms—specifically in *intentional* terms. The Time Lords (incorrectly) perceive Kevin as threatening consistency, and so coordinate their interventions with his attempts so as to stop him.

Now, we have seen that we need not suppose Time Lords or such like to exist, but the only alternative seems to be that systematic *uncaused* correlations are responsible for Kevin's failure(s). And, Horwich claims, these UCs are highly improbable. Hence P2.

3.5. Objections: Typicality

One complaint that has been raised regarding PGA-1 is not so much with the argument itself, but with its ability to establish anything about time travel in general. One might allow that PGA-1 shows *K* is improbable, but deny that *K* represents a typical TTS. Sider[9] argues that PGA-1 cannot constrain the probability of TTSs which do not involve agents attempting to go back and change the past. TTSs in which people just go back and live quiet, consistent lives, or where some non-human object goes back, are untouched by the argument. If this is true, PGA-1 is of very limited scope and significance.

[9] Sider, 2002, p.119

I will not address this objection here, for it really requires some way to speak broadly about what the consequences of different sorts of TTSs are, which itself requires a general framework not introduced until §6. In that section, I argue that even if a person is not intent on changing the past, they will still tend to act in ways problematic for time travel.

3.6. Objections: Improbability of Uncaused Correlations

Other objections focus on P2. Some objections to do with the vagueness of the probabilistic claim being made[10] are hopefully avoided here by my having committed to a certain interpretation of probability early, but other more specific ones remain.

One influential objection to PGA-1, raised by Smith[11], disputes that UCs are always inherently improbable. Smith admits that we do normally consider UCs to be improbable. If the White House head of security, upon hearing of plot to kill the president, proposed to rely upon the assassins' slipping on banana peels, or his gun jamming, to ensure the president's safety during the parade, we would politely point out that this is not a reliable safety policy. According to Smith, though, the improbability of these foiling events is not based on any deep physical fact about UCs. It is rather a purely inductive inference from the fact that almost all correlations we tend to observe have some explanative causal antecedent. Yes, we sometimes see UCs, as when I happen to bump into a friend at the supermarket. But these are rare, and thus induction warrants low degrees of belief in the occurrence of any particular UC.

Smith goes on:

> Now in so far as we can, in general, trust inductive inferences, we can trust that in contexts similar to those in which we have observed certain phenomena to occur only rarely, the phenomena in question will not occur very often. We can, however, draw no conclusions concerning the frequency with which the phenomena will occur in contexts unlike those in which we made the observations. (Smith, 1997, p.369-70)

[10] Smeenk and Wüthrich, 2011, p.9-11
[11] Smith, 1997, p.367-71

For example, consider the probability of A: *that the Sun will go out tomorrow*. Inductive inference ensures that given just that the Sun has never gone out so long as anybody has been observing it (discounting night time and eclipses), the probability of A is very low. But this information only constrains the probability so long as no *better* information is available. If a certain physicist claims that the Sun is running out of fuel, one cannot still simply appeal to the historical luminosity of the Sun to rule out A. That is not to say that the physicist cannot be challenged, and with information ultimately derived inductively, but this information needs to address why the lack of hydrogen does not support A, rather than just citing past matters of fact.

According to Smith, claiming that UCs are improbable even in the presence of time travel is like claiming A is improbable even when the Sun is running out of hydrogen. In both cases, the evidential context has been shifted by the presence of new and more powerful information. The past scarcity of UCs cannot ground probabilities in the context of time travel because this context brings its own reason to expect UCs to occur: to foil attempts to change the past. Thus, P2 is not justified by the available evidence.

Sider also endorses this reasoning, saying that 'the present absence of coincidences does not seem to be projectible'[12], and thus that it cannot support general statements about time travel.

3.7. Ditching Uncaused Correlations

I agree that the only evidential support for the improbability of UCs seems to be largely inductive. But the reason for this is not particularly deep; it is just that *uncaused correlation* does not seem to capture any precise or clearly bounded set of phenomena. For a start, it is a mystery what exactly is to count as a correlation. All around us every day there are occurrences which happen to be correlated with respect to *some* feature. If I shake up a bag of rice there will inevitably be grains which end up pointing in the same direction. Is this a UC? Presumably not; the significance given to UCs in the above debate implies that the correlated features must belong to some special class. But nobody seems to have attempted to spell out what this class is.

[12] Sider, 2002, p.120

Thus, the best conception of UCs we have is that based off of the few canonical examples usually given: slipping on a banana peel, one's gun jamming, being struck by lightning. It seems hopeless to attempt to apply any lawlike generalisation to such a loose and heterogeneous class of phenomena. Smith and Sider's criticisms of Horwich thus do not really have much to do with time travel per se, but rather with the fact that PGA-1 relies centrally on general claims about a very under-defined sort of phenomena.

While it is possible that a precise definition of UCs in theoretical terms could justify the claim that they are generally improbable, this is not an efficient way of approaching the issue. It seems to me that the notion of UCs is introduced really as a way of making the argument more general; since on PGA-1 we need not say exactly what improbable event an arbitrary TTS entails, just that it is some *sequence* of UCs, we can conclude that the TTS is improbable.

In the following, I drop all reference to UCs (and coincidences) and pursue a different strategy for arguing that TTSs are in general improbable. TTSs' improbability arises not from their entailing any particular type of event, but from the fact that they tend to imply that events (of no particular kind) both do and do not occur. Despite this change in emphasis, though, I think the idea behind PGA-1 is fundamentally correct, and my own argument should be construed as really a reformulation of it.

4. Records, Dispositions, and Self-Subversion

In this section, after making precise the ideas which underlie the reasoning of PGA-1, I present and defend my improved version of Horwich's argument: PGA-2.

4.1. The Probe Scenario

Since the goal is to make as precise as possible the ideas behind PGA-1, we ought to switch to an example TTS which relies on clearer assumptions than the Kevin scenario, thus allowing us to more easily see exactly what the problem with it might be. This new example TTS is the *probe scenario*, which is the scenario PGA-2 specifically argues is improbable. It will be the focus of §4-5, after which we turn to more general scenarios.

The probe scenario (now taking the label *K*) is inspired by an example in Earman[13]. Suppose there exists a space ship, the *mothership*, which is able to launch a robotic probe. The probe has a radar scanner connected to an anti-collision safety circuit, designed to prevent it from launching so long as it detects any objects in the vicinity of the mothership. The probe scenario goes: such a probe travels back in time through a wormhole and arrives at the position of the mothership while the latter is preparing to launch the (same) probe into the wormhole.

Fundamentally, this is the same sort of scenario as Kevin's; a system goes back in time and is disposed to interfere with the process that led to the system's setting off in the first place. And, as before, something happens which prevents this interference. The simplicity of the probe scenario derives from the fact that there is really only one thing which could prevent the interference, which is that the safety malfunctions, hence allowing the launch despite the presence of the probe. It also eliminates any confusion which comes with talking in terms of human agency (which is why Earman originally introduced it).

Now, depending on what *K* says about the details of the safety and its propensity to malfunction, the probability of the launch will seem to vary. If it is an old, primitive design, the launch may seem not altogether improbable. If, on the other hand, it is a cutting edge, carefully designed, and finely

[13] Earman, 1972, p.231-2

engineered device, we would expect that the launch is extremely improbable. Indeed, the probability of the launch can seem to be made arbitrarily low if enough extra details regarding the safety are added to the scenario. We will return to whether this supposed improbability can be made sense of later. First, we need to break down exactly how K bears on the probability of the launch.

4.2. Records

A scenario K comes with an associated set of *records*. Records are members of $\aleph_K$, meaning they are properties assigned by the scenario to the system at the latest point on its trajectory C_K. Records are distinguished by their being properties a system tends to acquire from interacting with its environment, and as such they tend to provide good evidence about events which take place in the vicinity of earlier parts of the system's trajectory C_K. (Note that, as with any mention of evidence in this essay, records provide evidence only *relative* to the background information E.) I will talk of records ambiguously between being properties of systems and of scenarios, with the second usage derivative on the first.

Human memory signatures and written accounts provide excellent examples of records, firstly because their complexity allows them to carry a lot of information, and secondly because there are strict principles for how certain events get recorded in the structure of these systems' properties (though in the case of memory these principles are known only implicitly). This means these records are easily 'decoded'. Other examples might be a fossil, whose form provides good evidence of the existence of certain creatures; or a scraped car, whose damaged state indicates it was in an accident. Clearly, what counts as a record will be fairly loosely defined, but I think the notion is a clear one.

In the case of the probe, its position and velocity are records of its past positions (assuming it is moving inertially). As is any writing on it. Its design is a very good record, implying many things about many aspects of human technological prowess. Its AI will have stored records of radar reports and internal diagnostics. All in all, given only modest background information, the probe has a huge variety of records, providing evidence of various strengths about a huge variety of different things. One proposition for which the records plausibly provide stronger evidence than anything else is Q: *that the probe is launched*, since this is seemingly the only way the probe could have got to where it is.

4.3. Dispositions

A scenario K also comes with an associated set of *dispositions*, also members of $\aleph_K$. Dispositions are distinguished by their being properties of a system which tend to lead to that system's interacting with its environment at later times in certain ways. As such, they provide reasonably good evidence about what events will happen in the vicinity of p_K^1 (recall: the latest point of C_K, but not necessarily the latest point of the system's maximal trajectory). I will say that a system has a *disposition to X*, which means that its dispositions imply that X is the case. Use of the term *disposition* is not supposed to explicitly suggest any link with the metaphysical idea of dispositions[14], though the notions are probably related. I shall also speak of dispositions as properties of both systems and scenarios.

Some examples. A murderous person has a disposition to commit murder. An armed bomb has a disposition to explode. A cue ball traveling towards another ball has a disposition to deflect that ball. The scope for dispositions is somewhat more limited than that of records, and the evidence they provide tends to be less strong. This is because it is in general much harder to predict what will happen to a system on the basis of its current state than what *has* happened to it. I will not speculate as to *why* this is; I simply assume this record-disposition asymmetry in evidential strength from here on.

Nevertheless, dispositions can provide important evidence about the future of a system. In the case of the probe, the probe has a disposition to continue on an inertial trajectory. It has a disposition to continue logging its radar scans. And it has a disposition to be detected by its past self, and hence to trigger the latter's safety, preventing the launch. That is, K is disposed to $\neg Q$. Throughout the following I will assume that everything a scenario implies (except what the scenario explicitly says) is implied by some combination of just its records and dispositions. I shall also, when discussing what propositions a scenario does or does not imply, refer to these propositions as *outcomes* (for records as well as dispositions).

[14] See Choi and Fara, 2018 for more on dispositions.

4.4. Self-Subversion

Recall that the idea behind PGA-1 was to show that K implied some improbable proposition, and must therefore itself be improbable. Following through this reasoning seems to require finding some proposition A such that $P(A|KE) \approx 1$ (i.e. K implies it) and $P(A|E) \ll 1$. Can this be done for the probe scenario?

Let's examine what outcomes A such that $P(A|KE) \approx 1$ might themselves be improbable. Some salient ones are: i) that the probe was built by humans, ii) that it passed through the wormhole, iii) that the probe launches (Q). (i) is positively probable; so far as we know (and hence so far as the background information E is concerned), all advanced technology like the probe was built by humans. (ii) is a little more subtle. The probe's travelling through the wormhole plausibly requires certain special conditions, as discussed in §2.5, and so one might think that (ii) *is* improbable relative to E, if E includes the physics we think we know. However, I argued that our understanding of the physics is simply not developed enough to ground judgements like this. Thus, for the purposes of this investigation, we cannot assume (ii) is in itself improbable.

Clearly, though, (iii) is the thing we want to say is improbable. After all, if the launch requires the malfunction of the very reliable safety, then it seems the launch must be improbable. But this is jumping the gun. PGA-1, in its current form, requires that Q be improbable just relative to E. But the background information on its own implies nothing about whether the launch occurs; that the launch does not occur is only implied by E *plus* the fact that there is an object in the vicinity of the mothership. So far as just E is concerned, the launch is neither probable nor improbable, since it depends too much on local matters of fact. (One *could* argue that Q is improbable on just E, but they would need to present some general reason weighing against such a launch taking place — i.e. one that does not rely on the specifics of the probe scenario.)

So K seems to imply nothing straightforwardly improbable in itself. But is this just a peculiarity of the probe scenario? Of course, there *are* some TTSs which imply outcomes which are straightforwardly improbable. For instance, a version of the probe scenario in which the probe is carrying a sealed vial of radon-222 (half-life 4 days), and *not one* of whose atoms has decayed during

the (three day long) journey, would be implying something extremely improbable. Similarly, if the Kevin scenario had included that Kevin was a three-time lottery winner prior to his trip, it would also imply something improbable. Consequently, we would be justified in disbelieving that either of these scenarios occur.

Clearly though, the elements of these scenarios which imply improbable outcomes do not really have anything to do with the fact that they are *time travel* scenarios. They are no different from any scenario in which some improbable contrivance is shoehorned in. The question is whether TTSs have any special tendency to imply improbable outcomes, *qua* being TTSs.

Returning to K, let's try to spell out the exact sense in which the launch *is* improbable. As has been argued, it is not the case that Q is improbable *in itself* (given just E). Nor is it the case that Q is improbable given K; even though K is disposed to $\neg Q$ (by virtue of the probe's position being such to tend to trigger the safety), K's records completely overshadow this disposition. As stated in §4.3, records tend to outweigh dispositions, and K is a typical example of this. If one were to come across the probe as it floats back towards the mothership, given its trajectory, the writing on it, the AI records, one would be overwhelmingly certain that the launch takes place, even despite the strong disposition the probe has to prevent the launch.

This observation does provide a clue of how to proceed, though. The idea at the core of the probe scenario is that the launch is improbable not relative to K, but to a certain logical consequence of K. This consequence is D: *that there is an object near the mothership*. D more or less implies exactly what the dispositions of K do, but without having the attached records implying Q. Hence, it seems highly plausible that $P(Q|DE) \ll 1$, for all D implies is that the safety will be triggered, and so the launch cancelled.

This suggests a new way of spelling out Horwich's rationale for this specific example, captured in the following argument (which I contrast with PGA-1 below).

The Probabilistic Grandfather Argument 2 (PGA-2):

P1: $P(Q|KE) \approx 1$

P2: $P(D|QKE) \approx 1$

P3: $P(Q|DE) \ll 1$

P4: $P(K|E) \leq P(Q|DE)/[P(Q|KE)P(D|QKE)]$ (See Appendix §8.2)

C1: $P(K|E) \ll 1$ (From P1 - 4)

Before discussing the merit of this argument, allow me to introduce a new piece of terminology.

Self-Subversion: A scenario K *self-subverts* iff there exist two propositions A and B such that $P(A|BE) \ll P(A|KE)P(B|AKE)$.

Self-subversion will be discussed in more depth below.

It is a simple consequence of the definition of conditional probability (see §8.2) that if a scenario self-subverts, then the scenario is very improbable. Thus, PGA-2 can be viewed fundamentally as an argument that K self-subverts.

What is the justification of P1 – 3? P1 has been discussed in §4.2; the records of K mean that it strongly implies that Q. P2 follows from the fact that D is entailed by K. Of course, in general the probability of D given any proposition entailing Q (in this case QKE) would be expected to be very low, due to the safety's reliability meaning that the launch's occurring strongly implies that no object was nearby. However, D is still consistent with Q (since malfunction is always possible), and so, that K entails D ensures that P2 is true. P2 could also be true even if D were not entailed by K, so long as it was implied strongly enough to outweigh the negative evidence provided by Q.

The important premise is P3, which basically represents the idea that presence of an object near the mothership, absent any other specific considerations, provides excellent evidence that the probe will not launch. This is, after all, the whole point of the safety.

I will discuss criticisms of P3 in the next section, but first I think it makes sense to explain why the term *self-subversion* was chosen. The point of introducing PGA-2 and self-subversion is to capture the idea that the probe scenario implies by virtue of some of its features that a certain thing happens, while simultaneously implying by virtue of its other features that that same thing does not happen. This is basically the probabilistic version of arguing that K entails a contradiction; a self-subverting scenario has two elements which are *nearly* contradictory, as it were.

The thought behind PGA-2 is that TTSs have a special tendency to self-subvert, lacked by other kinds of scenario, because in TTSs there are outcomes constrained by *both* the scenario's records and dispositions. If it happens that the records and dispositions imply very different things regarding these outcomes, the TTS self-subverts. In contrast, the records and dispositions of non-time travel scenarios cannot conflict in this way, because the lack of TTTs locks records to past outcomes and dispositions to future outcomes. Thus, TTSs tend to be much less probable than other scenarios. This is the *idea*. Whether it is true in general is a question which will have to wait until §6. In the mean-time, there are some more details to attend to.

4.5. The Context Objection

PGA-2 has a distinct advantage over PGA-1 in that it is not in terms of UCs, and so the issue of what UCs are and in what contexts they are improbable is completely sidestepped. Instead, it identifies a certain outcome Q and argues directly that it is improbable. Now, I have not tried to establish the improbability of Q from any general or theoretical considerations. I am simply assuming that to most readers it is obvious that $P(Q|DE) \ll 1$ (P3), given the facts about the safety mechanism. The idea is that this is true when no time travel is involved, and since it seems to involve no reference to time travel, it should also be true when time travel *is* involved.

Still, it is possible to make basically the same criticism of PGA-2 as Smith does for PGA-1 (§3.6) (though I think its appeal is diminished). While it seems intuitive that $P(Q|DE) \ll 1$, it may be argued, this is not supported by the evidence; we simply do not know how probable the launch is given D *in general*. All that we can say is that the launch is improbable given D *and* given that the launch takes

place in a *non-time travel context* ($\tilde{T}$-context). This is because all the evidence we have which weighs upon the probability of Q given D was gathered in a $\tilde{T}$-context.

The idea of a context, as with the original criticism, is a vague one. Let us just suppose that something happens in a $\tilde{T}$-context iff nothing to do with time travel is involved in the event, and that something happens in a *time travel context* (T-context) iff it does not happen in a $\tilde{T}$-context.

The objection may be made precise in the following way. Where T is the proposition that the launch occurs in a T-context and $\tilde{T}$ is the proposition that it occurs in a $\tilde{T}$-context,

$$P(Q|DE) = P(QT|DE) + P\left(Q\tilde{T}|DE\right) = P(Q|TDE)P(T|DE) + P\left(Q|\tilde{T}DE\right)P(\tilde{T}|DE)$$

The above follow from standard probabilistic identities. If we apply the fact that $P\left(Q|\tilde{T}DE\right) \ll 1$, which is the intuitive and true proposition which P3 is (according to this objection) masquerading as, then

$$P(Q|DE) \approx P(Q|TDE)P(T|DE)$$

Since we are allowing ourselves no assumptions about the probability of time travel *simpliciter*, we cannot ascertain the value of $P(T|DE)$. And since we have not gathered any data in time-travel contexts, we cannot ascertain the value of $P(Q|TDE)$. Hence, goes the objection, P3 is completely unjustified. Furthermore, if we substitute $\tilde{T}D$ in for D in PGA-2, we get P3, but we lose P2 because $\{\tilde{T}, K\}$ is (presumably) inconsistent. Call this the *context objection*.

I find little intuitive appeal in the context objection, but given its similarity to the highly popular objection of Smith to PGA-1, it is worth addressing to the fullest possible extent. My aim in the next section is to directly refute the context objection. It is my hope that the sort of reasoning displayed in this refutation of the context objection will put to rest more general doubts about the validity of my probabilistic approach to time travel.

5. Chance

This section sees the introduction of chance and the presentation of the chance argument as a means of defeating the context objection once and for all.

5.1. Preliminaries

We will now be dealing with *two* kinds of probabilistic quantities: the kind which I have been using up to now; and objective chance. From here, I will continue to refer to the former as *probability*, while the latter will be *chance*. *Probable* shall mean *having high probability* and *likely* shall mean *having high chance*. The converse for *improbable* and *unlikely*.

Specific positive characterisation of chance will have to wait, but what chiefly distinguishes it from probability is that it is neither intrinsically normative nor related to degree of belief (though it relates to the latter *via* the principal principle (discussed later)). Rather, it is supposed to be a more or less physical feature of the world and systems within it, akin to mass or length. It only directly has to do with persons or their beliefs insofar as these exhibit chancy physical characteristics.

In this section I will be working towards presenting an argument for P3. The argument basically goes: the launch has a low chance of occurrence in light of D; the chance of Q is independent of the T-context; and thus the probability of Q given D is low. Call this the *chance argument* (CA). Despite the simplicity of the CA, spelling it out mathematically requires a very precise formal definition of chance. Consequently, little progress can be made until we have in hand such a formal theory of chance.

§5.2-4 will be discussions of two independently plausible theories of chance (one of which is mine) which could be used for the CA. I should emphasise that by *theory* I here mean a formal apparatus for assigning chances to propositions and generating chance statements. I will not discuss interpretations of what these chance statements actually mean until §5.5.

I should also say that while I bring chance into this essay primarily because I think it is the reason P3 is (and appears to be) true, there are also interesting insights to be had about chance more generally and how it needs to be adjusted to deal with time travel.

5.2. Temporal Theory

In his work on probability[15], Lewis takes chance to be given by a function from proposition-time pairs to $[0,1]$. This theory of chance, which I call the *Lewisian Temporal Theory (TMP)*, is probably the most well-known, which is why it will be discussed first. We will see that, while elegant, it is too simplistic to adequately model how chance works in the presence of CTCs, and hence is unable to be used in the CA. This is a rather subtle result, however, and there is a somewhat more immediate inadequacy in Lewis' original theory, which is that, as with his definition of time travel[16], it relies on time being absolute. So the first job here is to put TMP in terms of spacetime geometry.

I propose we take chance on TMP as given by $Ch(A,\sigma)$, a function from proposition-spacelike hypersurface (SH) pairs to $[0,1]$ (where σ is the SH). The SHs take the place of the times, and so it would be natural to take these SHs to be time slices (i.e. edgeless SHs), since these are usually considered to be the spacetime analogues of absolute times. However, chronology violating spacetimes tend to have complications when it comes to time slices. No CVS is entirely foliable into time slices, and some (e.g. Gödel spacetime) have no time slices at all[17]. Thus, since we need chance to be defined for these CVSs in order to run the CA at all, it makes sense to allow SHs with edges into Ch's domain.

Before expanding on this, let me define the *apposite region* Γ_A of a proposition A. Γ_A is the minimal closed region of spacetime encompassing all of the events which make any difference to the truth of A. One might define apposite regions in terms of regions of possible worlds, as Lewis does[18] (though using different terminology), but I shall leave them as relatively intuitive entities. Some non-physical propositions may have the null set as their apposite region, while particularly general propositions may have the entire spacetime. However, we will be dealing here only with propositions where the apposite region plausibly has relatively small size and clear boundaries (e.g. Wellington wins the battle of Waterloo).

[15] See 1980, 1994
[16] See Lewis, 1976
[17] Earman, 1995, p.168
[18] 1980, p.272-3

Returning to chance, we might allow arbitrary SHs into the domain of Ch. Consider, though, that the underlying idea of TMP is that chance is a part of physical reality, and as such is subject to something like causation. The occurrence of some event may cause a change in the chance of A in the same way an event may change the length of a particular steel rod. For instance, the chance of my completing my jigsaw puzzle on a particular day may have been high throughout that day, until an unfortunate gust of wind scatters the pieces across the dining room table, after which it may be very low. In this case, the gust has intervened to cause this change in the chance.

Certain SHs only have the chances defined on them they do by virtue of their relations with surrounding physical goings on. Thus, if we commit to chance's being constrained by locality, it makes sense to limit the SHs $\{\sigma\}$ for which the chance of a proposition A is defined to be only those such that $\Gamma_A \cap I^-(\sigma) \neq \emptyset$ or $\sigma \cup I^-(\Gamma_A) \neq \emptyset$.[19] This basically means that those SHs consisting of points completely causally isolated (spacelike separated) from Γ_A will not have a chance for A defined on them. (If causation is nonlocal this constraint may have to be dropped.) This will not be hugely important in subsequent discussion, but it serves to highlight the dynamical nature of chance on TMP, which *will* be important.

With the basic stuff out of the way, we can begin to see the overall idea of TMP. On it, a proposition does not simply have a low or a high chance *simpliciter*, but only relative to a certain time (shorthand for SH). The chance of a proposition may vary over time, and when considering what events might affect the chance of a proposition, one has to consider at what times these events will actually have their effect felt.

So the rough idea with the probe scenario is that, while the chance of the launch was initially not particularly low (for the same reason $P(Q|E) \ll 1$ is unwarranted), the arrival y of the probe near the mothership (i.e. the event of the probe's reaching p_K^1) makes the chance of the launch plunge to a tiny value (due to the probe's disposition to trigger the safety).

[19] I am using *chronological past/future I* above rather than *causal past/future J* because using the latter adds complexity without adding anything to the discussion. One would imagine, though, that light signals would also be relevant to chance, so if I were here presenting TMP as a full theory, I would broaden the constraint appropriately.

5.3. The Fixing Problem

Showing exactly how this drop in chance is supposed to occur, and how it justifies P3, would require going into TMP in a lot more depth. This could be done; it can be shown that the Principal Principle (more on this later), together with the probe's ability to drop Q's chance, entails P3. However, this would be a waste of time, since there is a fundamental problem with the supposition that the probe's arrival e lowers Q's chance. This problem arises ultimately from the principle of historical fixing.

Lewis claims that the chance of a proposition A at all times strictly *after* its apposite region (though in his own terminology) is either 1 if A is true, or 0 if A is false[20]. I call this (the principle of) *historical fixing*. Translated into terms of spacetime regions, it goes: for any SH σ, if $\Gamma_A \subseteq I^-(\sigma)$, $Ch(Q, \sigma) = 1$ if Q is true, and $Ch(Q, \sigma) = 0$ if Q is false.

At a technical level, this time asymmetry is introduced by Lewis to account for the fact that one can know with pretty much certainty after a proposition A's apposite region Γ_A whether or not A is true (due, one supposes, to the high evidential power of records (see §4.2-3)), and hence personal probabilities in the future of this region tend to be 1 or 0. Since Lewis claims that in this case one should still set one's personal probability equal to chance (see §5.7), keeping persons rational requires that the chance of A be 1 or 0 after Γ_A.

Whatever the precise reason, though, historical fixing is part of TMP as presented by Lewis, and it has important consequences. Recall that the basic idea we are pursuing is that the arrival y of the probe drops the chance of Q to very low. y's dropping of the chance necessarily consists of a difference between the chance of Q on subsequent SHs. Suppose there are two SHs σ_1 and σ_2, where y is after σ_1 and before σ_2. We suppose that $Ch(Q, \sigma_1)$ is not low. What must be claimed is that $Ch(Q, \sigma_2)$ *is* low, since this is the outcome of y's having dropped the chance of Q.

Now, $\Gamma_Q \subseteq I^-(p_K^1)$, since the probe would have been present for its own launch. And, in general, all outcomes A implied by a scenario K's records will be such that $\Gamma_A \subseteq I^-(p_K^1)$, since records provide evidence by being the result of some causal interaction. *And* $p_K^1 \in I^-(\sigma_2)$, since σ_2 is after y. Thus, $\Gamma_Q \subseteq I^-(\sigma_2)$ and, by historical fixing, $Ch(Q, \sigma_2)$ is 1 or 0. This is basically a mathematical expression

[20] 1980, p.278

of the idea that the only SHs on which the probe will have a causal influence are those where the chance of Q is already fixed as 1 or 0 (since, from the probe's perspective, Q has already happened). The upshot of this is that TMP cannot be used for the CA, since depending on whether σ is before or after y, $Ch(Q, \sigma)$ will either be independent of D, or $Ch(Q, \sigma) = 1$ (or 0, but this is pretty much ruled out by K). Either way, even without precise argument, it is clear that this lends no support to the claim that $P(Q|DE) \ll 1$.

One might attempt to salvage the situation by denying historical fixing, or interpreting it differently from the way I have done. After all, while historical fixing is very central to Lewis' whole approach, it is logically independent from the rest of TMP. Thus, it is not impossible to endorse a version of TMP in which historical fixing is altered or removed so as to avoid leading to the fixing problem. I am not aware of any such version, though, and cannot see how one would really work. It just about makes sense that SHs prior to Γ_A encode certain causal determiners of A's chance, but it is extremely unclear what could possibly ground changes in its chance *after* Γ_A. Historical fixing is attractive in most contexts because it allows us to avoid answering this bizarre question.

So what does all this mean for P3? TMP is in general a pretty attractive theory; it is simple, and it captures a lot about how the idea of chance tends to be used. Hence, TMP's failure to support P3 might suggest that chance is simply unable to justify the claim, and may be unsuitable for application to time travel scenarios in general.

This would be jumping the gun, though, for there is good reason to expect that TMP will give incorrect results when nonstandard spacetimes are involved. This is because the idea implicit in TMP is that the chance of a proposition A is determined collectively by the events in the spacetime. In simple spacetimes, a single hypersurface will be able to capture all that causally influences events in Γ_A, but in more complex spacetimes this need not be the case. See the diagram.

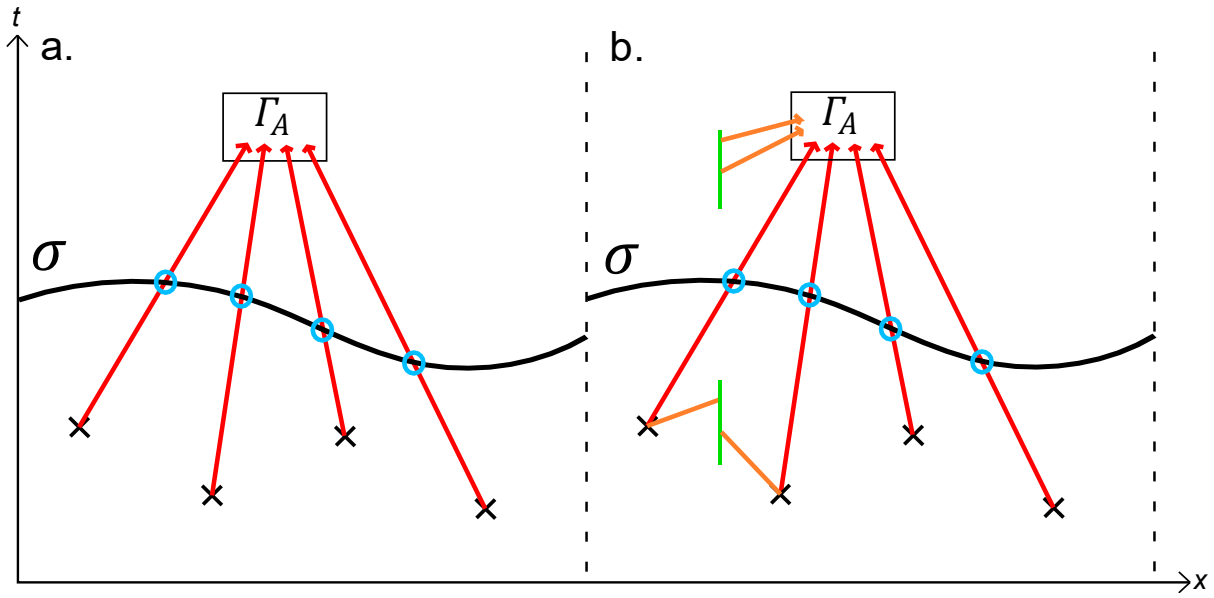

Figure 2 (colour required): We see two similar situations in which a group of events influences the chance of A via causal chains (red/orange) from each of the them to Γ_A. In (a), the spacetime has planar topology, so each causal chain must cross the HS σ. Thus, every causal influence on A up to that point is reflected in the events on σ, and so it makes sense to define chance on the HS. However, in (b) there is a wormhole (green) which allows causal chains to reach Γ_A without passing through σ. In this instance, the events on σ do not reflect *everything* influencing A. It seems plausible that the (b) situation will be common in non-planar spacetimes, and so a particular SHs will not provide a good reflection of the factors determining a proposition's chance.

The above is only a sketch, but I think it is enough to demonstrate that TMP is constructed specifically for topologically simple spacetimes (for which it functions very well), and that it is not really a suitable way of spelling out how chance works at the more general level.

5.4. Setup Theory

The above suggests that a theory of chance which takes finer account of how the causal nexus determines chances is required, not just for the task of justifying P3, but in general. An account which achieves this to a much greater extent than TMP is Setup Theory (SPT). While certain elements of the SPT are far from new, I have had to invent all of the formal and technical elements just to get a notion of chance precise enough for use in the CA.

On SPT, chance is given by $Ch(A, B)$, a function from proposition(-proposition) pairs to [0,1]. Ch is not in general defined for all proposition pairs. A *System* ϕ (I capitalise the *S* to distinguish this from *systems* as introduced in §2.1) is a consistent set of propositions having form $[Ch(A, B) = x]$. Chance is *defined on* ϕ *for* $\langle A, B\rangle$ iff $[Ch(A, B) = x] \in \phi$ for some value of x. Consistency requires that if $[Ch(A, B) = x] \in \phi$ and $[Ch(A, B) = y] \in \phi$, then $x = y$.

There is some System Φ such that, $[Ch(A, B) = x]$ is true iff $[Ch(Q, S) = x] \in \Phi$. In other words, the actual values of chance are given by Φ. Chance is *defined* (simpliciter) *for* $\langle Q, S\rangle$ iff $[Ch(A, B) = x] \in \Phi$ for some value of x. For those pairs $\langle A, B\rangle$ for which chance is defined, A is the *outcome* and B is the *setup*.

A System ϕ is not in general deductively closed. However, there is a set $\bar{\phi} \subseteq \phi$ called the *core of* ϕ. $\bar{\phi} \cup L$ *is* deductively closed, where *L* is the set of all natural laws ($\bar{\phi} \cup L$ is generally *not* a System). We can pick out a subset of $\bar{\phi}$, its *axioms set* $\bar{\bar{\phi}}$, which contains all and only propositions in $\bar{\phi}$ which are not entailed by any subset of $\bar{\phi} \cup L$ of which they are not a member[21]. Note that the core and axiom set are themselves Systems. So the core consists of just the axioms and all the propositions which are entailed lawfully by them (i.e. entailed by $\bar{\bar{\phi}} \cup L$).

I cannot give a full account of how lawful entailment between chance statements works, but one clear case would be that if $[Ch(A, B) = x] \in \bar{\phi}$ and $[A \rightarrow A']$, $[B \rightarrow B'] \in L$ ($\rightarrow$ being the material conditional), then $[Ch(A', B') = x] \in \bar{\phi}$. There will of course be more complex relations of entailment, but these will have to remain implicit here.

The propositions in ϕ but not in $\bar{\phi}$ are part of the System because they involve information irrelevant to chance. C is *irrelevant on* ϕ *to* A iff for no pair $\langle A, B\rangle$ on which chance is defined for $\bar{\phi}$ does $B \vDash C$ (deductive entailment). This is basically to say that information is irrelevant to A iff it does not follow from the axioms and laws that it bears upon the chance of A.

A final claim. If C is irrelevant on ϕ to A, then $[Ch(A, B) = x] \in \phi$ iff $[Ch(A, BC) = x] \in \phi$. This is to say, adding irrelevant information to a setup does not change the setup's effect on chance (for

[21] Since $\bar{\phi}$ is not by itself deductively closed, these are not true axioms of $\bar{\phi}$, but I will still call them such.

a particular outcome), and adding irrelevant information to a proposition will not make it a setup for *A* if it is not already one.

Obviously, this is a lot of formal information to take in. The basic idea of SPT can be expressed relatively simply, though. Chance fundamentally enters the world through a set of axioms about certain phenomena. It then gets spread to a more general set of areas by virtue of how these phenomena interact with the rest of the world in line with the natural laws. In this way, the axioms and the laws pick out the elements of the world which fundamentally exhibit chance. Then, chance gets expanded out to as large a range of phenomena as possible (including phenomena with irrelevant elements), but applies only to those phenomena with some element which exhibits chance on the basis of the axioms and laws. The value of chance will be strictly fixed by the nature of these elements.

Here is a toy example. Suppose that it is an axiom of Φ (the true System) that the chance of rolling a six (outcome) on a roll of a D6 die (setup) is 1/6. Then suppose that my rolling a six on my next turn would cause me to win the snakes and ladders game on that turn, and that its being my turn causes me to roll the D6. Presuming this causal fact is captured in the natural laws (a debate I will not even begin to address), this means that the axiom and the laws entail that chance of my winning the game (outcome) on its reaching my turn (setup) is 1/6. In this way the 'basic' chances propagate out via laws into a wide range of areas.

Now let us suppose, plausibly I think, that none of the axioms of Φ together with the laws entails a chance of rolling a six (outcome) on its being a Tuesday (setup). The axioms and laws (let us suppose, for now) simply do not make any connection anywhere along the line between die rolls and days of the week. Thus, *its being Tuesday* is irrelevant to the outcome, which means that the chance of rolling a six on a roll of a D6 die on a Tuesday is just the same as on rolling a D6: 1/6. This seems to fit well with how we tend to make assignments of chance; irrelevant information can be imported in, but it makes no difference to the value of the chance.

What about non-irrelevant (relevant) information? For instance, suppose the die is weighted towards six. It will not be the case that when the fact the die is weighted is added to the setup, the chance of the outcome remains the same. This means the weighting must be relevant, and so the change in chance must follow from the laws and axioms. But this seems perfectly reasonable. The weight

distribution of a die seems like exactly the sort of thing the laws would have something to say about. We will return to the details of what the axioms actually are in §5.9.

Hopefully this makes the basic way SPT works relatively clear. It is a more clunky theory than TMP. However, it does not make any reference to times, which means it is fundamentally neutral on the topic of time travel. That is, unlike TMP, there is no assumption built into the theory which undermines P3 just on the basis that time travel is involved. Based on this, I will proceed to give the CA in terms of SPT chance. Before proceeding, however, it is necessary to have an idea of what actually determines which System is the true System Φ.

5.5. The Best System Interpretation

In the last section I skipped over the question of what determines which System is the true System Φ. Justifying P3 will ultimately require an answer to this question, and the most plausible answer is I think supplied by the *Best System interpretation* (BSI) of SPT.

The basic idea of the BSI is that Φ is the *best* System: the System which has the best combination of the three competing virtues: i) *simplicity* of axioms, ii) *breadth* of application, and iii) *fit* with actual matter of fact. The BSI was pioneered by Lewis[22], and while he (naturally) introduces it in terms of his temporal theory of chance, the fundamentals of the BSI do not require TMP. Lewis also throws it in with his best system analysis of laws, construing the chance axioms to be a species of law of nature. Hoefer[23] drops this and many other of these conditions, presenting an interpretation of what it is to be a best System less focussed on objective probabilistic laws and more on general usability. Hoefer's account is also already more or less in terms of SPT, too. Thus, I base my version of BSI on his (adjusted to incorporate SPT as I present it). Let's see how it works.

For a start, simpler is better. Simplicity for ϕ is ultimately a property of the axiom set $\bar{\bar{\phi}}$, and can be understood, according to Hoefer, roughly in terms of the degree to which it provides elegant unification and user friendliness. Those of a more quantitive leaning like myself might prefer to construe

[22] Lewis, 1994
[23] Hoefer, 2007

simplicity in terms of some measure on the number of axioms (though presumably not the number itself, for $\bar{\bar{\phi}}$ may be infinite), or something along those lines; there is some freedom when it comes to interpretation. However, there seems little reason to doubt that axiom sets will vary non-trivially in their simplicity, even over a fairly wide set of construals of *simplicity*. Even without a precise definition in hand, it seems uncontroversial that, *ceteris paribus*, the more axioms a System has, the less simple it will be.

Next, broader is better. Breadth is, roughly, the number of types of phenomena the system applies to. (Hoefer actually follows Lewis in calling it *strength*, but *breadth* I think better captures what the former had in mind.) This can be understood in terms of the proposition pairs for which chance is defined on ϕ. The more different types of outcome the System gives chances for, the broader the System is. (Note that broadness should not be thought of in terms of the variety of *setups*, since, from §5.4, variety in setups is just proportional to variety in irrelevant information.)

Finally, more fit is better. Fit is ordinarily taken to be the chance which ϕ assigns to the totality of true outcomes for which it is defined. That is, more or less, the chance of the world being as it actually is according to ϕ. However, this runs into problems for systems about an infinite number of occurrent phenomena, which would seem to necessarily assign zero chance to the totality of these outcomes. Furthermore, with STP it is not clear whether or how ϕ assigns a chance to the appropriate 'maximal fact'. One might consider the maximal fact to be the conjunction of all (pairs of) facts for which chance is already defined, but there is no guarantee that if chance is defined on ϕ for $\langle A, B\rangle$ and for $\langle A', B'\rangle$, that it will be for $\langle AA', BB'\rangle$.

Thus, Hoefer opts to construe fit more along the lines of *typicality*. As with simplicity, there is much room for interpretation of what this means. But again, regardless of exactly how one defines fit at the global level, we can be quite sure that, *ceteris paribus*, the higher the chance a System gives some particular true outcome, the better its fit. Relatedly, and more importantly, the closer the chance assigned by ϕ to $\langle A, B\rangle$ to the (actual) relative frequency (RF) of co-occurrence of A-outcomes with B-setups, *ceteris paribus*, the better ϕ's fit. This follows from the fact that the chance of a certain RF is highest if the chance of each 'trial' is equal to the RF.

There is little doubt that the three virtues relied upon by the BSI are rather vague and underdefined. However, together they paint a clear picture of what sort of features Φ will have, and thus give a good indication of the kinds of things which determine what is relevant to the chance of a particular outcome and what value this chance will have.

(Note that, on the BSI, there being nontrivial (i.e. > 0 and < 1) SPT chances is compatible with *determinism*: the claim that a specification of an (appropriately inclusive) boundary condition and all the natural laws, will specify all (physical) matters of fact. This is important because a lot of the examples I use involve deterministic dynamics. The rationale for this compatibility is that while determinism entails that a System with only trivial chance assignments (i.e. a deterministic System) over some boundary condition can achieve *maximal* fit and breadth, it does not entail this is the best System. Indeed, it seems perfectly plausible that even maximal fit and breadth is not worth the complexity of specifying *every* part of the boundary condition, suggesting that the best System will have nontrivial chances. Loewer gives an extended defence of this idea.[24])

5.7. The Principal Principle

One final idea is required for the CA: the *Principal Principle* (PP). First introduced in Lewis[25], what the PP basically says is that a person ought to set their degree of belief in a proposition to the chance they believe that proposition to have, provided they have no better information. Many take the PP to be definitional of chance[26], and even those who do not tend to think it is an important truth. The PP will have to be slightly modified from the form in which it is originally presented in order to deal with the formalism of the SPT, but the idea is fundamentally the same. It provides the key link between probability and chance required for the CA. (Lewis actually presents several versions of the PP[27]. For simplicity, I will be dealing only with the first one. While some adjustment would be required to deal with the conditionalized chance of the later version, the important elements would be precisely the same.)

[24] Loewer, 2004
[25] Lewis, 1980
[26] e.g. Wallace, 2012, p.139-51
[27] 1980, 1994

Here is my updated version (all mentions of the PP from here refer to this version):

The SPT Principal Principle (PP): $P(A|BE) = Ch(A, B)$ so long as *i)* BE does not entail B' where $B' \vDash B$ and $Ch(A, B) \neq Ch(A, B')$, and *ii)* BE entails no inadmissible information regarding A.

Conditions (i) and (ii) basically together ensure the PP applies only when there is no better information than that represented by the chance available. Condition (i) ensures that no information from E would shift the chance of A, were it added to B. This essentially means that all the information relevant to (the chance of) A present in BE is present in B. For instance, take the example of where the weighted D6 is thrown. Even if the die is weighted, the chance of rolling a six on *rolling a D6* is still 1/6. But the PP does not here require us to have degree of belief 1/6 in rolling a six, given the D6 is rolled. This is because E contains the information that D6 is weighted towards six. Since the chance of rolling a six on rolling a D6 *which is weighted towards six* is (we suppose) greater than 1/6, (i) is not satisfied for $\langle A, B\rangle$, and the PP does not apply. (Of course, provided no *more* relevant information is in E, the PP *does* imply that we should set our degree of belief in rolling a six to whatever the chance of six on a *weighted* D6 is thrown is.)

Condition (ii) is found in Lewis' original version. If (i) ensures that no better *chance* information is available, (ii) ensures no better *non-chance* information is available. The paradigm piece of inadmissible information for A is A itself; the PP does not apply if one already believes that A actually is the case, since if one believes A, one believes it regardless of its chance. There is other, less direct inadmissible information, though. Lewis tries to characterise in general what information is *admissible*, saying that it tends to include historical and hypothetical information,[28] but we have seen in §5.3 that this leads to historical fixing. It is also difficult to see how historical information is picked out in SPT, where neither chance nor probability is intrinsically temporal. Hoefer[29] points out these difficulties, and opts to not attempt a reduction of admissible information to other kinds of information, instead just

[28] 1980, p.272-6
[29] p.552-5

taking it to be information whose only evidential importance to A is through A's chance. I follow Hoefer in this characterisation.

5.8. The Chance Argument

We now have the apparatus required to spell out precisely how considerations of chance imply P3, via the chance argument.

The Chance Argument (CA):

P1: $Ch(Q, D) \ll 1$

P2: $P(Q|DE) = Ch(Q, D)$ (From PP)

C1: $P(Q|DE) \ll 1$ (From P1 - 2)

There is much to be said in the defence of both premises. P2 has slightly fewer complications, so let's begin with that. As indicated, it follows from the PP. To check this, we need to consider whether the two conditions in the PP are satisfied for P2.

The first, that DE does not entail a stronger setup than D on which Q has a different chance, seems securely the case. By the design of the probe scenario, D covers everything which is relevant to the chance of that particular launch because the only thing specified by the example which has any impact on the launch is the detection of a nearby object. Sure, E contains a whole lot of extra ultimately relevant information, presumably including facts like that the probe will not launch if it has a fuel leak, or if its docking clamps jam, and a setup which included that any of these are the case would give a different value of chance. However, E does not entail that any of these actually are the case, nor does it give any indication of how likely they are. Thus, any stronger setup D' entailed by DE will presumably not differ from D in any respect relevant to chance, so will give the same value for the launch's chance.

The second condition, that DE entails no inadmissible information regarding Q, also seems securely the case. DE certainly does not *entail* Q or $\neg Q$, for it consists of just general information about the world and the statement that *some* object is near the mothership (not specifying that it is the probe). Nor does DE seem to even *imply* Q or $\neg Q$ in any way not plausibly attributable to its bearing on chance.

An example of a piece of information which *would* imply (but not entail) Q *not via chance* is F: that the probe bay was used in subsequent weeks for food storage. This would suggest that it was not occupied by the probe, thus implying that the launch took place. F gives some relatively specific information about events after the launch is scheduled, which results in its being inadmissible. Now, while D does in a sense give information about events *after* the launch, it is so non-specific that it cannot reasonably be called inadmissible. Thus, both conditions of the PP are satisfied for Q and D, and P2 thus follows.

Now for P1, which says that the chance of the probe launching (outcome) on its having detected a nearby object (setup) is very low. The first thing to note is that a necessary condition for P1 is that chance is actually defined for $\langle Q, D\rangle$. This seems pretty unproblematic, though. If chance is defined for anything, one would think it would be defined for the sorts of highly mechanical and regular phenomena which occur in electronic systems like that of the safety. And since the launch is causally determined (i.e. by natural laws) to occur or not by whether the safety malfunctions or not, based on the contents of §5.4 the launch itself will have a chance assigned to it on an object's being detected. In terms of the best System Φ, if the core $\overline{\Phi}$ assigns chances to the various phenomena concerned in the functioning of the safety, it will assign chances for those setups causally determining and those outcomes causally determined by those phenomena – i.e. the detection of an object and the launching of the probe, respectively.

Furthermore, these same sorts of considerations strongly suggest that $Ch(Q, D)$ will have a small value. After all, if a system is specifically designed to reliably produce a definite and precise outcome on certain antecedent conditions, it seems pretty probable that whatever physical principles lead to this reliability will be such to entail the chance of failure to produce the outcome is small. This rough idea will be spelled out more thoroughly in the next section when we consider the role of irrelevance in the chance argument.

First, though, we must return to the context objection of §4.5. Recall that this consists of pointing out that $P(Q|DE) \approx P(Q|DTE)P(T|DE)$, where T is the proposition that the launch (and the arrival of the probe) occur in a T-context. The objection then goes that our current evidence, which was

gathered in a $\tilde{T}$-context, does not support any claims about the value of $P(Q|DTE)$. Since the CA purports to constrain the value of $P(Q|DTE)$ just from available evidence, it is inconsistent with the context objection, basically entailing that current evidence is projectable into a T-context.

Assuming one accepts P2 (which, I have argued one ought to), then the context objection basically becomes the claim that the evidence does not support the claim that $Ch(Q, D) \ll 1$; it only supports that $Ch(Q, D\tilde{T}) \ll 1$. The CA relies on a rejection of this, on the basis that chances are invariant between T-contexts and $\tilde{T}$-contexts. The STP allows this to be stated in a precise form:

$$Ch(Q, DT) = Ch(Q, D\tilde{T}) = Ch(Q, D)$$

and to be given a precise explanation: that T and $\tilde{T}$ are both irrelevant to Q. I call this claim, which will be defended in the next section, *context irrelevance*. It entails that if $Ch(Q, D\tilde{T}) \ll 1$ then $Ch(Q, D) \ll 1$ and thus that P3 is true. If context irrelevance can be shown to be true (or at least plausible), then there seems little reason left to reject P3, and thus to deny that K self-subverts.

5.9. Context Irrelevance

My argument that T-context is irrelevant is based on the observation that a proposition's relevance to the chance of an outcome generally depends directly on its *dynamical* importance to it. If something does not play some clear role in the dynamical process which gives rise to an outcome, it is probably not relevant to that outcome's chance of being true.

To see this, we need to consider what sort of axioms $\overline{\overline{\Phi}}$ is liable to contain, given the way the world works and the sort of phenomena occurrent in it. For example, take the general phenomena of *die rolls*. Now, there are a fair few kinds off dice which get rolled: D20, D12, down to the favourite D6 and even D4. The RF of the outcome *die shows 1* (called: 1) to the setup *a Dn is rolled* (called: Dn) for these dice are (very nearly) 1/20, 1/12, 1/6, and 1/4, respectively. These are RFs of extremely long and stable sequences, and so it is relatively safe to assume that a large amount of fit is gained by a System's assigning chances equal to these RFs to the respective relevant outcome-setup pairs, and thus that this is what Φ does. Now, this might just be because the appropriate chance statements $Ch(1|D6) = 1/6$,

$Ch(1|D20) = 1/20$, etc. are all members of $\overline{\overline{\Phi}}$. But this is a very non-simple way to achieve this fit; a System assigning chances to rolls directly requires as many axioms as there are kinds of dice (not just now, but ever, across all of space and time).

A far simpler way is, if possible, to have some single axiom which lawfully entails that $Ch(1|D6) = 1/6$, $Ch(1|D20) = 1/20$, etc. This results in precisely the same fit, but for a massively simpler axiom set. Is there such an axiom? Well, in the case of dice, the fine dependence on initial conditions means that the phase space representing a die's initial velocity and position (angular *and* linear, and relative to the landing surface) divides into roughly equal size regions corresponding to the die's rolling a particular number. How many different-number regions there are, and what parts of the phase space they occupy will depend on the number of sides the die in question has. But whether it is a D4 or a D100, it will be the case that each possible number will have an equal phase volume.

Thus, the only axiom needed to get the fit is α: $Ch(N, I) = V_N/V_0$ (where N: *die shows number N*, I: *initial state* $\in V_0$, and V_N is the volume corresponding to N). α entails the chances of all numbers for all possible (fair) dice, and this is basically because it introduces chance only at a very general level. It then relies upon the *natural laws* to determine via a whole complex set of dynamic equations which volumes correspond to which numbers for different dice. This can be seen as the System exploiting the complexity already present in the system of natural laws to supply the complexity in chance statements, thus keeping $\overline{\overline{\Phi}}$ relatively simple. The simplest and hence probably best System will be the one which most fully exploits ultra-simplifying axioms like α which are in terms of things the laws of nature heavily constrain, like physical dynamics.

Now the link to relevance begins to come into focus. If we assume that Φ will tend to have axioms which apply chance just to very general physical phenomena, and that particular outcome-setup pairs have chances defined by virtue of their being certain special instances of these phenomena, then a proposition will in general be relevant to an outcome only to the extent that it describes something physically operative in the way the outcome comes about. For example, *that it is Tuesday* is irrelevant to *rolling a 1* because the day of the week plays no dynamical role in the process which leads to a certain

side facing up. This is in contrast to the number of sides, or the centre of mass of the die, since both represent features operative in the dynamics of a rolling die.

These principles can be extended to more complex phenomena. Suppose there is a particular chance of my getting married before turning 23 (on some setup or other). I have no idea exactly what sorts of processes are involved in determining when (or if) I marry, but I expect they are probably hugely complex and diverse. Thus, it makes little sense for me to try and deduce what this chance is in terms of the operation of these processes. However, what I *can* be sure of is that if some fact is definitely not involved in the process, then it will be irrelevant to the chance of the outcome. Hence, regardless of anything else, I am quite sure that my star sign is irrelevant to my chance of marrying before turning 23, because I know of no plausible mechanism (except possibly a firm belief in astrology) by which the precise time of one's birth influences when one marries decades down the line.

Parallel reasoning can be applied to the T-context and Q. There simply does not seem to be any way in which the time travel dynamically influences the launch of the probe. Sure, the probe *has* time travelled, and *it* dynamically influences the launch. But it influences it by virtue of being a physical object in a certain location at a certain time; not by its having time travelled. Given the way that the safety (and hence the launch) works, the influence on the launch of the probe is identical to that of some random space rock floating by. This is what was meant by saying that D (which makes no mention of time travel) captures *all* the dispositions dynamically relevant to the launch.

And this is the case more generally. Simply being 'in the context' of time travel—being related in some salient way to it—is irrelevant to the chance of an outcome so long as there is not some specific physical way in which the time travel itself is supposed to influence the process leading to the outcome. This is, as I have argued, because the best System is going to be one which applies chances along the boundaries in the world carved out by the natural laws, and since most natural laws seem to be dynamical, by the dynamics.

Thus, $Ch(Q, DT) = Ch(Q, D)$. And since the above all applies equally to $\tilde{T}$, $Ch(Q, D\tilde{T}) = Ch(Q, D)$. Since I do not think anyone would dispute that $Ch(Q, D\tilde{T}) \ll 1$, P1 of the CA follows.

This concludes my chance-based argument against the context objection, and hence for P3 of PGA-2. While I talked explicitly only about chance in the probe scenario, I hope it is clear that my points apply more generally; wherever there is chance involved in a TTS, there is no reason not to rely on evidence collected in a $\tilde{T}$-context so long as the evidence represents the involvement of familiar physical processes in the scenario. Even when chance is not clearly involved (for whatever reason), the considerations of this section I think undermine the idea that generally time travel context should really have any importance when reasoning about TTSs.

6. Self-Subverting Scenarios

In this section, I discuss what types of time travel scenarios tend most to self-subvert, and in doing so address the generality of PGA-2 (i.e. whether it works only for a narrow, contrived set of scenarios).

6.1. Typicality Revisited

The aim of this final section is to establish some general principles about what sorts of TTSs tend to self-subvert, and what the features of these TTSs are which make them do so. Showing that a scenario K self-subverts basically consists of finding two propositions A and B such that $P(A|BE) \ll P(A|KE)P(B|AKE)$. In less precise but more familiar terms, it involves finding a proposition A which is implied to have occurred by the scenario, and a proposition B which is implied by the scenario (plus the occurrence of A), and which implies A does not occur.

So far, the discussion has been centred on the probe scenario. By design, there was only one thing clearly implied by the scenario, which was that the probe was launched. Hence this would have to be A. And, also by design, there was only one thing which would stop the launch: there being an object near the mothership. But it so happened that the scenario entails that this is so, because it describes the probe as coming back to its starting point. And hence, as I have argued at length, the probe scenario self-subverts.

Clearly, though, the details of the probe scenario are rather carefully chosen so that the time travel impacts on events in just the right way to produce self-subversion. It is improbable that any realistic ship would have such a simplistic safety mechanism, and it is even more improbable that the probe would end up returning to the mothership, rather than heading off into some other region of space (though still having time travelled). It is tempting to conclude from this that self-subversion only occurs for very contrived scenarios, where things are set up in exactly such a way for time travel to cause problems, and perhaps once more realistic TSSs are considered, we will fail to find any self-subversion.

This is reminiscent of the criticism of Horwich alluded to in §3.5: that his argument relied on the unrealistic assumption that time travellers would generally try to change the past (e.g. by murdering their ancestors). Given this, it makes sense to start by returning to the Kevin scenario and seeing if there

is any way to show that it self-subverts without having to introduce any unrealistic or contrived elements.

6.2. The Kevin Scenario: Redux

Recall the Kevin scenario. Kevin travels from 2019 Oxford to 1929 Portsmouth. Let's now drop the assumption, part of the scenario as given in §3.1, that Kevin intends to kill his grandfather. Thus the scenario is stripped of any clear contrivances; we may suppose that Kevin simply travels back to 1929 with no particular set of intentions. His properties include only typical properties of a 21st century British man: he wears modern clothes, carries a smartphone, remembers an assortment of facts about history, etc. Now, the question is: does this apparently more typical version of the scenario self-subvert?

First, what are the records of the scenario, and what do they imply? Kevin's mere existence implies some very broad things, like that all human life is not annihilated as a result of the Cuban missile crisis. His memories of learning about the 1939 invasion of Poland and the 1969 moon landings (given the background assumption that memories tend to be veridical) imply strongly that these events occur. The nature of his clothes implies quite a few things about changes in fashion and advances in polymer manufacture up to 2019, and the design of his phone implies much about 21st century electronics and telecommunications. His body itself has records implying there is a human with precisely the same DNA as his for a good period leading up to 2019; and of there being two other humans with very similar DNA to his existing during an earlier period, and so on. Thus, given only modest background information, a typical human like Kevin carries a vast amount of information about the world in his or her past.

Second, is Kevin disposed to contradict anything implied by the scenario's records? This, of course, depends in large part on his intentions. We are not supposing he wants to kill his grandfather, but there are still plenty of other problematic ways in which he might act. If he desires wealth, he might try to exploit his knowledge of the future to make a fortune at the bookies. If he desires acclaim, he might tell the world of his success at time travel and wow them with his technology and stories of the future. If he is a particularly moral individual, his memories of the evils of Nazism might motivate him

to try to assassinate Hitler prior to his rise to power. On its own, that someone with the specific knowledge Kevin has attempts to bring about any one of these things would imply that it does indeed come about.

And, of course, this itself would imply that many outcomes implied by the records are false. If someone became a billionaire by repeated (apparently) lucky bets, or became famous for time travelling, or if Hitler never rose to power, these would be things that Kevin would remember having occurred. Hence, that he does not remember them implies that they do not happen. If we take proposition A to be that *there is not a famous 1930s time traveller, nobody builds a fortune off of betting, Hitler rises to power, etc*. and proposition B to be *Kevin tries to become a famous time traveller, Kevin tries to build a fortune, etc*. then it seems reasonably clear that $P(A|BE) \ll P(A|KE)P(B|AKE)$ and that the scenario self-subverts.

One *can* argue against this conclusion. Perhaps Kevin is particularly ignorant of history, or takes no phone or clothes or anything which provides much of a record of events between 1929 and 2019. Or perhaps he is a particularly inept individual, who would be unable to impact the world of 1929 much even with his privileged knowledge of the future. Or perhaps he simply has no desire to do anything of note and lives a quiet, parochial life, perhaps out of fear of producing some cataclysmic time travel paradox, or because he believes the past cannot be changed. However, these all seem to me like contrivances of their own. If Kevin is just an *average* person, one would not expect him to exhibit any of these peculiarities, so it seems safe to say that the scenario tends to self-subvert.

The logic of the above objection is suggestive, though, basically exploiting the fact that the less either the records or the dispositions of the scenario imply about the period 1929-2019, the less clearly the scenario seems to self-subvert. It will be useful to spell this idea out in a more general form.

6.3. Evidential Overlap

First, let's assume that certain general features of scenarios affect the strength of their records/dispositions regarding the occurrence of certain kinds of events at certain times and places. The *strength* of a scenario's records/dispositions for a certain event type and spacetime region is, roughly,

the degree of evidential support the records/dispositions lend to events of this kind happening in this region. For instance, the probe scenario has strong records that a probe launch occurs in its local spacetime, and strong dispositions that no such launch occurs in this same region.

Now take the set of *event kinds* (characterised only roughly) Δ. For each $e \in \Delta$, a scenario K defines (at least heuristically) two scalar fields on spacetime: φ_R^e and φ_D^e, taking non-negative values at all points. They represent the scenario's strength of records and dispositions, respectively, for event kinds over certain regions of spacetime. The higher the average value of the field over a region, the greater the strength of records/dispositions for that event kind occurring in that region. Summing the φ_R^e or φ_D^e (appropriately normalised) over values of e gives another field, which represents roughly the general capacity of the scenario to record/influence events at certain regions of space. These fields are $\Psi_R(\boldsymbol{x}) = \sum_e \alpha_R^e \varphi_R^e(\boldsymbol{x})$ and $\Psi_D(\boldsymbol{x}) = \sum_e \alpha_D^e \varphi_D^e(\boldsymbol{x})$, where α_R^e and α_D^e represent weightings on event kinds (since presumably, on any perspective, some event kinds are more important to take account of than others). $\Psi_R(\boldsymbol{x})$ is the scenario's *record evidential density* and $\Psi_D(\boldsymbol{x})$ is its *disposition evidential density*. For brevity, I call them its *memory density* and its *influence density*. These together represent a system's ability to remember and influence events throughout spacetime (according to a particular scenario K).

Now we come to evidential overlap. The evidential overlap (*overlap* from here) a scenario has is $\theta = \int \Psi_R(\boldsymbol{x})\Psi_D(\boldsymbol{x})d\boldsymbol{x}$, where the integral is over the entire spacetime manifold. The higher the value of θ, the greater the extent to which the records and the dispositions provide evidence about the same subject matter (i.e. same sorts of events at the same time and place). *Ceteris paribus*, the greater the overlap, the more a scenario will tend to self-subvert, because greater overlap gives greater scope for there being outcomes A_i implied by the records whose negation is implied by the dispositions. If there are enough of these outcomes, then there will be an outcome $A = \bigwedge_i A_i$ (such as in §6.2) and an outcome B (capturing just K's dispositions) such that $P(A|BE) \ll P(A|KE)P(B|AKE)$.

6.4. Features Affecting Overlap

The point of presenting overlap in this formal manner is that it allows a clear sight of what different features of a scenario it is affected by. These features may be split into two: *innate capacities* and *spacetime displacement*.

Innate capacities are members of $\aleph_K$ tending to affect the values of $|\Psi_R(\boldsymbol{x})|$ and $|\Psi_D(\boldsymbol{x})|$, where $|\Psi_i(\boldsymbol{x})| \equiv \int \Psi_i(\boldsymbol{x})d\boldsymbol{x}$. The stronger a system's *memory capacity*, the greater $|\Psi_R(\boldsymbol{x})|$, and the stronger a system's *influencing capacity*, the greater $|\Psi_D(\boldsymbol{x})|$. Clearly, *ceteris paribus*, the greater either $|\Psi_R(\boldsymbol{x})|$ or $|\Psi_D(\boldsymbol{x})|$, the greater the overlap θ.

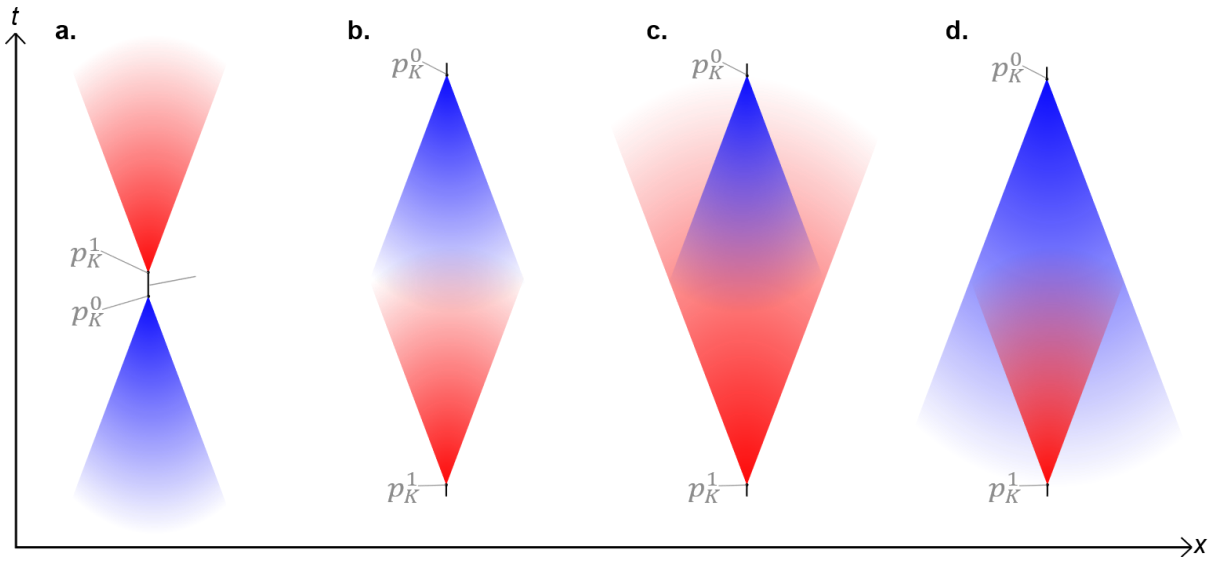

Figure 3 (colour required): Memory density Ψ_R (blue) and influencing density Ψ_D (red) for a series of scenarios. The more opaque a region, the higher the density there. (a) shows a typical non-time travel scenario in which overlap is zero, because Ψ_R is strictly about the past and Ψ_D about the future. (b) shows a TTS where there is only a small amount of overlap. (c) and (d) show similar scenarios but where $\aleph_K$ is altered to increase influencing and memory capacity, respectively. In both cases, overlap is clearly much greater than for (b). Note the shapes of the densities: I have assumed that the system may remember (influence) only events in the past (future) light cone of p_K^0 (p_K^1).

We will see some examples shortly of what sorts of properties might constitute a system's innate capacities.

The other factor affecting overlap is *spacetime displacement*: the magnitude of the interval between p_K^0 and p_K^1. In general, memory and influencing density will fall off at greater temporal and spatial distances from p_K^0 and p_K^1 respectively. This is because more distant events will tend to affect the system less strongly, making their leaving any records, or the system's influencing them, more difficult. As a result, *ceteris paribus*, the larger the (timelike) interval between p_K^0 and p_K^1, the less overlap a scenario has. If the interval is spacelike, there is no overlap.

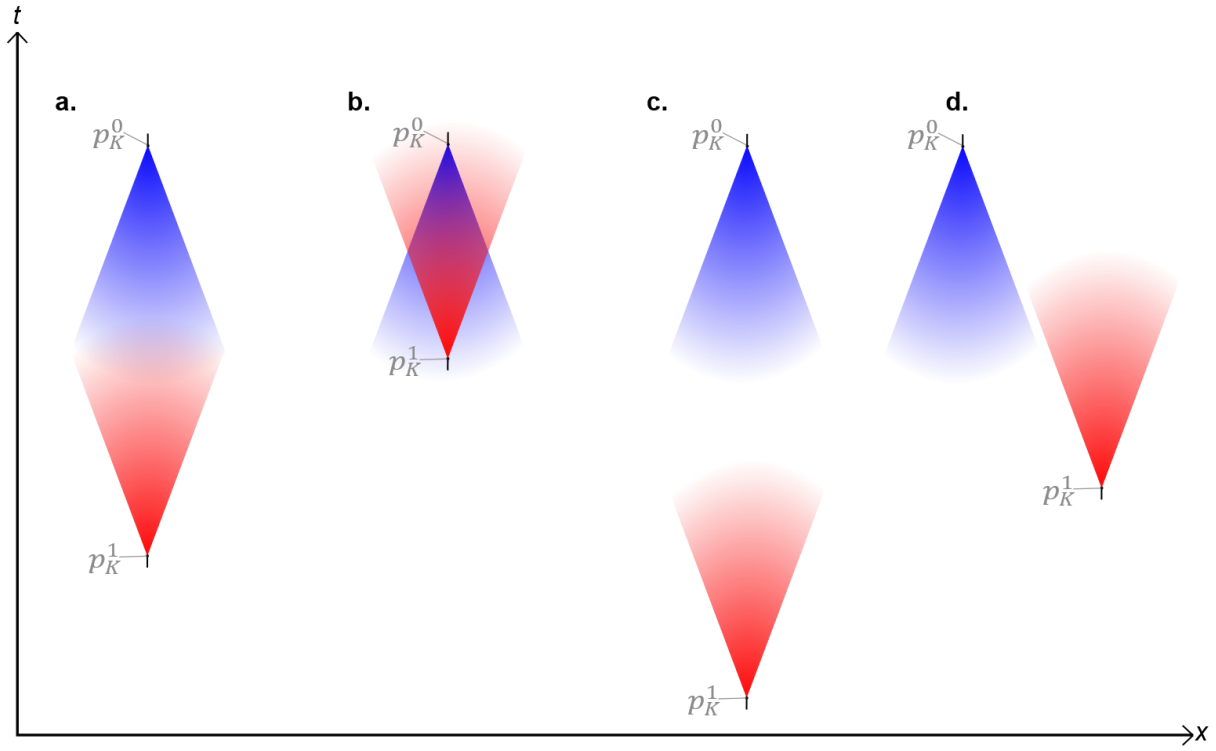

Figure 4 (colour required): Here we see four scenarios where the memory and influencing capacities are precisely the same (represented by the densities having the same size/profile in each case), but where the spacetime displacements result in differing overlaps for the scenarios. Note that the densities being zero outside the associated light cones means that for scenarios like (d), where the interval between p_K^0 and p_K^1 is spacelike, overlap is necessarily zero.

Memory capacity, influencing capacity, and spacetime displacement together go a long way to determining a scenario's overlap, and hence its tendency to self-subvert. Little can be said to constrain spacetime displacement, since what displacements are possible is a question of pure physics. However, what *can* be done is to establish the memory and influencing capacities of various realistic systems,

apply considerations of self-subversion, and thus deduce a rough lower bound on the spacetime displacements for scenarios involving these systems. This is the objective of the next three sections.

6.5. Simple Systems

It makes sense to first consider the innate capacities of very simple systems: ones with in general very few properties. One such system is the particle, whose properties tend to be more or less limited to position, momentum, spin, and some others. It is from this scanty set from which all records and dispositions of particle-scenarios must be drawn. Let's start with records, concentrating on position and momentum. By having its momentum and (hence) position altered by interactions with an external force field (possibly as a result of the presence of other particles), the particle carries information about the values of this field at points through which it has recently travelled. However, the lifespan of this information is presumably short, since the position and momentum will constantly be being altered by further interaction with the field(s).

The limited information carrying capacity of the particle system is a direct result of the small number of degrees of freedom (DoFs) it has (this is what makes it a simple system). While these DoFs may interact readily with the environment, generating information-carrying correlations with values of the field, the amount of information is limited by the number of DoFs. Plus, this information will only last as long as it tends to take for some new interaction to alter significantly a lot of the DoFs, which will presumably be short as there are so few of them. The details will depend upon the scenario; if the particle passes out of the field, or if it is very weakly interacting, then its records may last longer. But, in general, the memory capacities of simple systems like the particle are weak.

The situation is precisely similar for influencing capacity. The dispositions of the particle are its position and momentum, and by generating a force field of its own the particle will be able to influence other objects. But ascertaining what its influence will amount to is very difficult because (as mentioned) the particle will probably be constantly acted upon itself by fields. Consequently, the particle's influencing capacity is also poor.

This all adds up to mean that, unless a particle-TTS involves an extremely small spacetime displacement, there will be almost no overlap, and the TTS will not self-subvert.

6.6. Energetic Systems

Suppose a scenario has as its system a star. The memory of the star is, like the particle, very weak. While the star has a great number of DoFs (by virtue of being a very large object), it is so dense and isolated from other stars or planets that these DoFs do not really ever get correlated with the environment. Furthermore, the high temperatures mean that any degrees of freedom which do get so correlated will quickly have their information washed out by the strong and frequent internal interactions.

However, the massive energy of the star means it has a very strong, if imprecise, influencing capacity. It can have a whole system form around it; it can determine life to emerge within this system; and it can wipe this life out. It may be visible from thousands of light years away. And if it is massive enough, when it explodes it can destroy all objects within a huge radius.

Now, while the star may have very few records, one thing its records *do* imply is that the star has existed for (in general) a billion or so years. So any scenario in which a very massive star travels back to its own vicinity in the past and then hangs around until it explodes will without a doubt self-subvert. For the star's disposition to explode implies the destruction of its younger self, but its existence implies its younger self survives. From this it can be inferred that if the star is to time travel, it must either be too small to explode, or it must travel to a new spatial location. This is an instance of the general fact that even scenarios with very weak records may self-subvert if they involve a system with a strong enough influencing capacity, and influencing capacity is roughly proportional to energy.

6.7. Intelligent Systems

In §6.2 I discussed whether the Kevin scenario self-subverts, concluding that it does because Kevin would probably try to do certain things which he does not remember, and would remember had they actually happened. This is an instance of a more general trend, that scenarios involving intelligent

systems readily self-subvert, due to these systems tending to have both very strong memory capacities and very strong influencing capacities.

Intelligent systems like Kevin will plausibly in general have their own innate memory, but also will also likely produce and carry with them large numbers of records (both intentionally and unintentionally). This means that a scenario which specifies such a system will tend to imply a large number of outcomes over a fairly wide region of spacetime.

One can also expect intelligent systems to be able to manipulate their environment with a great degree of freedom and precision. Depending on the level of technology involved, the system may be able to do anything from lighting a forest on fire to carving its name into the nearest celestial body. However technologically advanced or powerful the system is, though, it will have a influencing capacity far stronger than a physically similar but unintelligent system.

Not only this, but it seems probable that if an intelligent system does time travel, it will be aware of it and will behave differently as a result. Be this due to curiosity, self-service, or some alien motivation, it seems quite certain that a system accustomed to reacting in complex ways to its environment would not simply ignore that it had time travelled. In this respect intelligent systems may be unique: the difference between TTSs and other scenarios involving the same intelligent system is not just that Ψ_R and Ψ_D overlap more in the former, as with most systems, but also that Ψ_D may itself be altered (and probably expanded due to the system's having knowledge of the future) by the occurrence of time travel.

Thus one can expect intelligent-system-TTSs to in general have large overlap and to self-subvert readily. The strength of the memory and influencing capacities means that the only such TTSs not extremely improbable are ones where the spacetime interval is very large. For instance, while Kevin can make a problematic stir in 1929, if he instead travelled back to 2M B.C, there is not really anything Kevin can do which would show up in his own 2019 records, since most traces of his presence would erode away over the next 2M years. However, even a scenario with this displacement *could* self-subvert if the involved system had a strong enough influencing capacity (e.g. if Kevin brings a super-laser and burns his name into the moon).

7. Conclusion

This essay has been an attempt to place an upper bound on the probability of certain kinds of time travel, based on the degree to which they involve a conflict between the records and dispositions of the time-travelling system. I emphasise *upper bound*, because this conflict is far from the only factor constraining TTSs, not least of which is whether the required physical mechanisms for time travel actually exist.

In §6, I argued that, while the majority of systems could time travel without creating serious problems, intelligent systems of the kind often seen in sci-fi are very prone to leading to self-subversion. Even so, while these specific results are useful in the time travel debate, the main utility of this investigation has been its presenting a general method with which one can approach an arbitrary TTS and, by identifying its evidential overlap, place a solid constraint on its probability. I hope that the clarity offered by this way of doing things goes some way to clearing up the confusion often present in debates on time travel.

An obvious way to expand on the material covered here would be to consider the self-subversion tendencies of a wider and more detailed set of TTSs. Similarly, it would be helpful to have a discussion of how (SPT-)relevance considerations weigh on a broader range of phenomena, since this would provide a firmer basis for the claim that context is unimportant when reasoning about self-subversion. Indeed, it seems worth exploring and expanding SPT for its own sake, since it seems like an eminently plausible theory of chance, and one which I believe has not been presented before.

8. Appendix

8.1.

For any propositions A, B and background information E,

$$P(A|E) = P(AB|E)/P(B|KE)$$

$$\leq P(A|E)/P(B|KE)$$

where the last line follows because $P(AB|E) \leq P(A|E)$, necessarily.

8.2.

For any propositions A, B, C and background information E,

$$P(A|E) = \frac{P(AB|E)}{P(B|AE)}$$

$$= \frac{P(ABC|E)}{P(B|AE)P(C|ABE)}$$

$$\leq \frac{P(BC|E)}{P(B|AE)P(C|ABE)}$$

$$\leq \frac{P(B|CE)}{P(B|AE)P(C|ABE)}$$

For P4 of PGA-2 substitute in $A = K$, $B = Q$, $C = D$.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Bibliography

Earman, J. (1972). Implications of Causal Propagation Outside the Null-Cone. *Australasian Journal of Philosophy*, 50 (3). (pp.222-37).

Earman, J., (1995). *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*. New York: Oxford University Press.

Choi, S., and Fara, M. Dispositions, *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition), https://plato.stanford.edu/archives/fall2018/entries/dispositions/.

Hoefer, C. (2007). The Third Way on Objective Probability: A Skeptic's Guide to Objective Chance. *Mind*, 116 (2). (pp.549–596).

Horwich, P. (1987). *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge MA: MIT Press.

Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly*, 13. (pp.145–52).

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In Richard C. Jeffrey (Ed.) *Studies in Inductive Logic and Probability*, Vol II., Berkeley and Los Angeles: University of California Press. (pp.263-93).

Lewis, D. (1986). *Philosophical Papers: Volume II*, Oxford: Oxford University Press.

Lewis, D. (1994). Humean Supervenience Debugged. *Mind*, 103. (pp.473–490).

Loewer, B. (2004). David Lewis's Humean Theory of Objective Chance. *Philosophy of Science*, 71 (5). (pp.1115–1125).

Malament, D. (1985). Minimal Acceleration Requirements for Time Travel in Gödel Space-Time. *Journal of Mathematical Physics*, 26. (pp.774-777).

Malament, D. (2012). *Topics in the Foundations of General Relativity and Newtonian Gravitation Theory*. London: University of Chicago Press.

Smeenk, C., Wüthrich, C. (2011). Time travel and time machines. In Callender, C., (Ed.), *The Oxford Handbook of Philosophy of Time*, Oxford: Oxford University Press.

Wallace, D. (2012). *The Emergent Multiverse*. Oxford: University Press.