Causation and Time Reversal

Forthcoming in *British Journal for the Philosophy of Science*.
Latest version of paper: mattfarr.co.uk/files/theme/ctr.pdf

Matt Farr
University of Queensland
✉ mail@mattfarr.co.uk | mattfarr.co.uk

November 26, 2016

Abstract

What would it be for a process to happen backwards in time? Would such a process involve different causal relations? It is common to understand the time reversal invariance of a physical theory in causal terms, such that whatever can happen forwards in time (according to the theory) can also happen backwards in time. This has led many to hold that time reversal symmetry is incompatible with the asymmetry of cause and effect. This paper critiques the causal reading of time reversal. First, I argue that the causal reading requires time-reversal-related models to be understood as representing distinct possible worlds, and on such a reading causal relations are compatible with time reversal symmetry. Second, I argue that the former approach does however raise serious sceptical problems regarding the causal relations of paradigm causal processes, and as a consequence there are overwhelming reasons to prefer a non-causal reading of time reversal whereby time reversal leaves causal relations invariant. On the non-causal reading, time reversal symmetry poses no significant conceptual nor epistemological problems for causation.

1 Introduction

What would the world be like if run backwards in time? This question is ambiguous since it depends upon whether a 'backwards-in-time' world would involve an inversion of cause and effect. It is common to understand the invariance of a physical theory under *time reversal*—an operation that takes a motion to the temporally reversed motion—as entailing that whatever can happen for-

wards in time (according to the theory) can happen backwards in time, implying that causal relations are reversed under time reversal. On such a reading, the time reversal symmetry of fundamental physics appears incompatible with the asymmetry of cause and effect, and has consequently been taken by many to motivate eliminativism about causation in physics and at the 'fundamental level' more generally. This so-called *Directionality Argument* traces its roots back at least as far as Bertrand Russell's (1913) defence of causal scepticism. In what follows, I argue that such worries about time reversal are misplaced on two grounds. Firstly, a causal interpretation of time reversal (whereby time reversal inverts cause and effect) requires us to understand models related by a time reversal transformation as representing distinct possible worlds. On such a reading time reversal symmetry is compatible with the existence of directed causal relations since each such world preserves the asymmetry of cause and effect. However, I show that this approach does lead to major conceptual and epistemological problems regarding the direction of causation for the kinds of systems to which we typically assign unambiguous causal judgements. Secondly, and consequently, I demonstrate that there are overwhelming reasons to reject a causal interpretation of time reversal. Rather, causal relations should be understood to remain invariant under time reversal. On the preferred non-causal reading of time reversal, I show that time reversal symmetry poses no major conceptual nor epistemological problems for causation. Moreover, this reading fits naturally with popular accounts of causal discovery.

The paper is structured as follows. The rest of the introduction covers the paper's background by outlining the Directionality Argument and the concept of time reversal. Sec. 2 asks what time reversal reverses. I consider two rival philosophical accounts of time reversal: the *C theory*, according to which pairs of models related by a time reversal transformation represent a single possible world; and the *B theory*, according to which time-reversal-related models represent distinct possible worlds. I show that a causal reading of time reversal requires the *B* theory, but that the *C* theory is preferable on independent grounds. Sec. 3 asks whether time reversal reverses causal relations. I argue that time reversal should be interpreted non-causally, and defend the epistemology of causal direction this provides. Finally, sec. 4 asks whether time reversal symmetry is compatible with causation. I show that on both causal and non-causal readings of time reversal compatibilist accounts of causation and time reversal invariance are available. Sec. 5 considers consequences of the paper's conclusions.

1.1 The Directionality Argument

Philosophical consideration of causation typically concerns a relation, R, that holds between a pair of events, c and e, where $R(c, e)$ is read as 'c causes e'.[1] Such a relation is standardly taken to be *asymmetric*, such that if an event c is a cause of an event e, then e is not a cause of c:

$$R(c, e) \rightarrow \neg R(e, c). \tag{1}$$

Thus, for a pair of causally-related events a direction of causation can be defined insofar as one event causes the other and not *vice versa*. We can say that c is the cause and e the effect, and there is a fact of the matter as to which is which and in which direction the causal influence propagates. The causal relation is also assumed to be *time-asymmetric* insofar as causes temporally precede their effects. Thus, it follows from $R(c, e)$ that c is earlier than e:

$$R(c, e) \rightarrow E(c, e) \tag{2}$$

where E is the 'earlier than' relation.[2]

These features of causation are widely held to be incompatible with time symmetries of fundamental physics. In particular, time symmetry plays a central role in Bertrand Russell's (1913) case for causal eliminativism in (classical) physics. Russell cites time symmetric features of the law of gravitation (taken by Russell as an exemplar of physical laws) as incompatible with both the asymmetry and time asymmetry of causation. This *prima facie* incompatibility has been presented as an argument in the recent literature under the name 'the Directionality Argument',[3] which runs roughly as follows:

1. If the fundamental physical theories are time-symmetric then they are not causal.

2. The fundamental physical theories are time-symmetric.

3. Therefore, the fundamental physical theories are not causal.

[1] I assume the causal relation holds between pairs of events for reasons of simplicity. What I say can be extended to more complicated cases, such as where an effect has multiple causes, and *vice versa*, and also where the relation holds between type events or possible values of variables.

[2] Though these two features fall far short of a full 'folk theory' of causation, they suffice for the aims of this paper, which is to assess whether such an account of causation is compatible with time reversal symmetry.

[3] See Field (2003), Ney (2009), Frisch (2012) and Farr and Reutlinger (2013) for versions of the Directionality Argument.

The argument trades on the intuitive incompatibility of the (time) asymmetry of causation with the time symmetry of physical theories. The notion of 'time symmetry' is clearly central to the argument, however it is ambiguous. A theory can be thought to be time symmetric in at least two distinct senses: first, it can be invariant under a set of well-defined time reversal transformations; second, its dynamical laws can be of such a form that, relative to some given state of a system, they determine or give non-trivial probabilities for its possible past and future trajectories.[4] This second kind of time symmetry may be termed the *bidirectionality* of its laws or predictive algorithm.[5] The Directionality Argument has been discussed in terms of time reversal invariance by Field (2003, p. 436), Ney (2009, p. 747), Norton (2009, pp. 481–2) and Frisch (2012, p. 320).[6]

1.2 Time reversal

Time reversal may be understood in classical terms as a set of operations that reverse a physical motion. For example, the time reverse of a ball rolling from left to right is a ball of equal mass rolling with the same speed but from right to left. A theory is *invariant* under time reversal if and only if the time reverse of every motion allowed by the theory is also a motion allowed by the theory. This entails that if a theory (1) models some particular process x, and (2) is time reversal invariant, then it follows that the theory also models the time reverse of x. As such, a time reversal invariant theory can model any allowable process relative to either time direction. For convenience, call a pair of models related by a time reversal operation *TR-twins*. The contention of this paper is that the relationship between causation and time reversal importantly depends upon whether or not one takes TR-twins to represent distinct possible states of affairs.

Intuitively, time reversal inverts the time order of a sequence of states of a system by taking the time coordinates from t to $-t$. However, in general time reversal also involves an operation on the instantaneous states of systems. For

[4] Farr and Reutlinger (2013) argue that Russell's discussion of time symmetry appears to refer not to time reversal invariance *per se* but rather to the bidirectionality of the law of gravitation, in that it nomically entails the past and future trajectories of a given state—Russell (1913, p. 15) holds that "[t]he law [of gravitation] makes no difference between past and future: the future 'determines' the past in exactly the same sense in which the past 'determines' the future."

[5] By 'predictive algorithm' I have in mind the Born Rule in quantum mechanics. In the case that such an algorithm is bidirectional, 'predictive algorithm' is a misnomer — such an algorithm would then also be *retro*dictive.

[6] Of these, only Norton explicitly endorses the claim that time reversal invariance is incompatible with the asymmetry and time asymmetry of causation. Frisch rejects such an argument. Field and Ney both make implicit reference to both time reversal invariance and the predictive/retrodictive symmetry of classical theories in discussing Russell's claim.

example, the standard time reversal transformation for Newtonian mechanics involves velocity reversal. Newtonian mechanics is intuitively time reversal invariant insofar as for any motion of particles allowed by the theory (assuming the elasticity of collisions, etc.), the time-reversed motion is also allowed by the theory, and so the theory admits of no irreversible processes. However, since the instantaneous state of a Newtonian system includes velocities, a time reversal operation that merely inverts the sequence of states fails to secure the time reversal invariance of the theory, since the velocities of particles must also be inverted in order for the new sequence of states to satisfy the equations of motion.[7] As such, we may understand time reversal as taking a sequence of states $S_i, \ldots, S_f$ to ${S_f}^*, \ldots, {S_i}^*$, where the '*' superscript denotes the time reversal operation on the instantaneous state. This feature of time reversal is common across physical theories.[8]

The action of time reversal upon properties of the instantaneous states of a system brings up two important points concerning the relationship between causation and time reversal of relevance to this paper. Firstly, switching a process for its TR-twin implies not only a passive coordinate transformation—i.e. a shift in perspective—, but also an active transformation upon physical quantities of the system. For example, Maudlin (2007, p. 119) holds that the need to apply time reversal to instantaneous states implies that "even for an instantaneous state, there is a fact about *how it is oriented with respect to the direction of time*" [emphasis in original].[9] On the contrary, in the next section I defend a fully passive interpretation of time reversal (the *C* theory), whereby TR-twins, despite potentially differing with respect to quantities of instantaneous states, nonetheless equivalently represent a single possible world. Secondly, the action of time reversal upon instantaneous states might be seen as an *ad hoc* device designed purely to secure the time reversal invariance of a theory.[10] This brings in

[7] In classical Hamiltonian mechanics, a state is given by the three-dimensional position and momentum values of the particles. Here, the momenta are vectorial properties — they are the product of velocity and mass. As such, everything I say about the direction of velocities can be translated to talk about the direction of momenta should the reader wish. This distinction makes no difference to the points made about time reversal and causation.

[8] For instance, the standard set of time reversal transformations in electrodynamics inverts the magnetic field, and in quantum mechanics inverts spin, etc. See Sachs (1987) for a detailed account of time reversal operators across physics.

[9] The idea of an instantaneous state being time directed has itself been taken to be conceptually problematic. Albert (2000, p. 18) rhetorically asks "[w]hat can it possibly mean for a single instantaneous physical situation to be happening "backward"?" Callender (2000, fn. 4) objects that "[i]t just does not make sense to *time-reverse* a truly *instantaneous* state of a system."

[10] For instance, Arntzenius and Greaves (2009, p. 563) note that "any theory, including ones that are (intuitively!) not time reversal invariant, can be made to come out 'time reversal invariant' if we

an important pragmatic constraint on the form of a theory's set of time reversal operations: the time reverse of some process as determined by this set of operations must be a reasonable candidate for how that process would 'appear' if 'viewed' relative to the opposite direction of time.[11] This constraint is sufficient for the purposes of this paper. The independent and complex issue of the status and justification of particular sets of time reversal transformations for different theories is outside the scope of the paper.

Independently of Russell's motivations, the relationship between time reversal invariance and causation is interesting for its own sake. In particular, it is unclear in what sense time reversal invariance could be incompatible with causation. As Frisch (2014, p. 119) notes, the incompatibility of time reversal invariance and causation is often assumed without further argument, as though the incompatibility were self-evident. Such a view is mistaken. Upon analysis, I argue that time reversal and causation have a more subtle and interesting relationship. In order for time reversal invariance of physical theories to bear upon the metaphysics of causation, we require a philosophical account of how states of affairs are transformed under time reversal. In the next section, I outline two such accounts: the *B* theory and the *C* theory. We shall see how these different theories motivate different accounts of how causal relations transform under time reversal. Importantly, I argue that on both accounts time reversal invariance and causation are compatible.

2 What does time reversal reverse?

With the preliminaries out of the way, we may now turn to the compatibility of time reversal invariance and causation. My contention is that this depends upon whether time reversal is understood as inverting causal relations. We can set out two different readings of time reversal:

Causal time reversal (CTR). Time reversal involves inverting causal relations, taking causes to effects and *vice versa*.

Non-causal time reversal (¬CTR). Time reversal does not invert causal relations; the distinction between cause and effect remains invariant under time reversal.

place no constraints on what counts as the 'time reversal operation' on instantaneous states."

[11] Insofar as time reversal operations may be applied to in-principle unobservable processes (such as the quantum mechanical evolution of a system between measurements), the idea of a time-reversed state 'appearing' a certain way or being 'viewed' backwards in time is a heuristic metaphor.

While CTR is often assumed in the literature,[12] I'll argue that ¬CTR is preferable. Interestingly this issue has not been directly addressed in discussions of the Directionality Argument. Despite this, its relevance is clear. If time reversal inverts causal relations, then we face the following problem: if the world is described by a time reversal invariant theory, then any possible way the world could be is describable by at least two models of the theory (TR-twins) that ascribe different causal relations to the world. Conversely, if time reversal does not invert causal relations, then there is no *prima facie* conceptual problem of causation for a time reversal invariant theory, since TR-twins can share the same causal structure. This brings up two central aims of the paper. Firstly, I demonstrate that assuming CTR, the exact problem time reversal invariance poses for causation depends upon one's preferred temporal metaphysics. Secondly, I defend ¬CTR over CTR, and in this way argue that the time reversal invariance of fundamental physical theory would not warrant eliminativism about causation. This section addresses the first aim and lays the groundwork for addressing the second.

2.1 The *B* and *C* theories of time

The issue of CTR *vs* ¬CTR is importantly interconnected with whether one reads time reversal as an active or passive transformation—i.e. whether TR-twins represent different possible worlds or are different descriptions of a single possible world. For instance, a fully passive reading of time reversal, whereby time reversal is understood as nothing more than a redescription of a process relative to the opposite direction of time, implies ¬CTR. Such a view is outlined by Hans Reichenbach:

> Since it is always possible to construct a converse description [of a process], positive and negative time supply *equivalent descriptions*, and it would be meaningless to ask which of the two descriptions is true. (Reichenbach, 1956, pp. 31-32; my emphasis)

Reichenbach's suggestion is offered within an analysis of the time reversibility of classical mechanics. In virtue of the lack of irreversible classical mechanical processes, Reichenbach notes that classical mechanics may describe any allowable process relative to *either* temporal direction, and hence it is a matter of con-

[12] For instance, to note a recent paper in *Nature Physics*: "Under time reversal [...], states should become effects and *vice versa*" (Oreshkov and Cerf, 2015, p. 3).

vention to hold that classical mechanics in any sense describes or governs processes in the future direction only. As such, forwards-in-time and backwards-in-time descriptions of processes are strictly equivalent. Reichenbach relates this issue to the conventionality of geometry, where different geometries may be used to equivalently model a single state of affairs. On this account, time reversal is a passive transformation within an equivalence class of models: for any process, p, classical mechanics offers TR-twins describing p relative to each time direction, and both models should be understood as picking out one and the same possible process.[13] A similar view of time reversal as offering different but equivalent descriptions is offered by the cosmologist Thomas Gold:

> [T]he description of our universe in the opposite sense of time [...] sounds very strange but it has no conflict with any laws of physics. [The] strange description is not describing another universe, or how it might be but isn't, but it is describing the very same thing. (Gold, 1966, p. 327)

Gold, unlike Reichenbach, is here discussing a backwards-in-time macroscopic description of the world, containing putatively irreversible processes (with assumed underlying reversible mechanics). In this case, the forwards and backwards descriptions differ in that the latter describes apparently improbable behaviour (e.g. the anti-thermodynamic reforming of broken wine glasses, unmixing of coffee and milk, etc.) due to the presence of highly-correlated variables. (I argue in sec. 3 that such cases help to motivate the position of Reichenbach and Gold.)

The key idea present in Reichenbach's and Gold's suggestions is that TR-twins offer distinct but equivalent descriptions of a single possible state of affairs. For convenience, I will refer to this view as a *C theory of time*.[14] This terminology is motivated by McTaggart's (1908) distinction between the *B* series and the *C* series. Whereas the *B* series orders events in terms of the time-directed relation '*earlier than*', the *C* series is concerned only with the undirected '*temporal betweenness*' ordering of events:

> [T]he *C* series, while it determines the *order*, does not determine the *direction*. If the *C* series runs M, N, O, P, then the *B* series [...] can

[13] In cases in which a model is its own time reverse, e.g. a stationary particle, the same model describes both 'forwards' and 'backwards' versions of the relevant process.

[14] The *C* theory of time is presented and defended by Farr (MS). The claim that TR-twins are equivalent descriptions of a single state of affairs is entailed by the *C* theory but not exhaustive of it. For the aims of the present paper, this claim is the relevant feature of the *C* theory.

> run either M, N, O, P (so that M is earliest and P latest) or else P, O, N, M (so that P is earliest and M latest). And there is nothing [...] in the *C* series [...] to determine which it will be. (McTaggart, 1908, p. 462, my emphasis.)

The distinction between *order* and *direction* is key to the distinction between the *B* and *C* series.[15] McTaggart's usage of these terms is similar to Reichenbach's[16] own use of these terms in delineating his position regarding time order in time-reversible physics.[17] The *C* series is contrasted by McTaggart with the *B* series in terms of its lack of directionality. The *C* series of a set of events does not determine their *B* series: any time ordering of events in terms of temporal betweenness is compatible with two directed time orderings in terms of the 'earlier than' relation. This shows the difference in structure between the *B* and *C* series.

The *B* and *C* *theories* give two different ontologies of temporal relations. On the *C* theory, there are no time-directed states of affairs, and as such, no two worlds may differ solely with respect to the arrangement of 'earlier than' relations. We may understand time reversal as taking one *B* series of events arranged from earlier to later to the inverse *B* series by reversing each 'earlier than' relation. Such a transformation preserves the *C* series of the events, since it leaves the temporal betweenness relations invariant. The adirectional ontology of time given by the *C* theory entails that the two different time-directed pictures given by TR-twins differ only at the level of description—both TR-twins refer to the same time-direction-independent facts—and so the *C* theory entails the Reichenbach/Gold passive interpretation of time reversal.[18] Conversely, on the *B* theory, there are time-directed states of affairs, and so it follows that, first, two worlds may differ solely with respect to the arrangement of 'earlier than' relations, and second, TR-twins describe distinct possible worlds. We can take these as necessary conditions for *B* theory that suffice to distinguish it from the *C* theory.[19]

[15] Although McTaggart (1908, 1927) consistently refers to the *C* series as 'nontemporal', this is due to precisely the same reasoning for which he takes the *B* series to be nontemporal, i.e. that neither series contains 'real' (*A* series) change—in neither series is there a division between past, present and future that changes. Farr (MS) argues that a *C* theory of *time* is defendable once we relax the assumption that time requires *A* series change.

[16] Max Black (1959) similarly distinguishes the 'order' and 'arrangement' of a series of events, claiming that only the former is observable and hence fundamental.

[17] See Reichenbach (1956, chs. 2–6).

[18] In particular, any quantities that differ between TR-twins (such as instantaneous velocity, spin, etc., as discussed on p. 5) can be considered descriptive artefacts that equally correspond to a single time-direction-independent (*C*-theoretic) state of affairs.

[19] This is a non-standard way of presenting the commitments of a *B* theory of time. This is due

2.2 Time reversal on the *C* theory

In introducing the *B* and *C* theories, my aims are twofold: first, to show that both theories offer compatibilist accounts of causation and time reversal invariance; second, to argue that the *C* theory offers the superior account of both the function of time reversal and of the epistemology of causal direction. Since I am both proposing and defending the *C* theory, it is important to guard against possible objections and misunderstandings of its treatment of time reversal.

2.2.1 The *C* theory doesn't require time reversal invariance

John Earman (1974) objects to the passive interpretation of time reversal entailed by the *C* theory here presented, holding that such an interpretation "is too powerful; for this conclusion [that time reversal amounts to a redescription of a single state of affairs] follows whether or not the laws of physics are time reversal invariant" (Earman, 1974, p. 27).[20] We can understand Earman's point by noting that the following two questions concern distinct, though related, issues:

1. Do TR-twins describe distinct possible worlds?

2. Is some particular theory time-reversal invariant?

The former question divides the *B* and *C* theories. The latter concerns an independent issue that might be taken to motivate either theory, though does not objectively favour either. While the latter is a broadly empirical question, the former is an *a priori* issue concerning the interpretation of time reversal that is conceptually independent of whether some particular theory is time reversal invariant. Moreover, it is the former that directly concerns the relationship between causation and time reversal. Earman's implication is that the independence of these two issues is a problem for the *C* theory: the interpretation of

to the fact that the *B* series is standardly presented in negative terms—in that it does not commit to the *A* series' properties of 'pastness', 'presentness' and 'futurity', nor an objective passage of time—rather than in positive terms. However, the *B* series is characterised by the inclusion of 'earlier than' relations that are not present in the *C* series. Farr (MS) argues that the standard presentation of the negative and not positive aspects of the *B* series is due to the historical prominence of the debate over temporal passage which separates the *A* series from the *B* and *C* series. The separate issue of the *directionality* of time, which separates the *B* and *C* series, has occupied far less literature.

[20] Earman's criticism here is specifically aimed at the passive interpretation of time reversal defended by Black (1962), but applies also to his other targets, Reichenbach and Gold. Black claims that it would follow from time reversal invariance of fundamental physics that 'earlier than' is a three-place relation (such that x is earlier than y only *relative to* some third term z—e.g. an observer, some process, etc.). Earman rightly notes that Black's conclusion is actually a consequence of the passive interpretation of time reversal, and follows regardless of the time reversal invariance.

time reversal it offers is independent of whether the relevant physics is time reversal invariant. Importantly, the time reversal invariance of fundamental physical theories is neither necessary nor sufficient for the *C* theory. However, this is not a problem for the *C* theory in itself; rather, it highlights that the *C* theory primarily concerns an *a priori* issue that in turn determines one's understanding of time-asymmetric phenomena.[21]

2.2.2 Time reversal *non*-invariance on the *C* theory

Earman's worry does however point to a more general problem: in reducing time reversal to a redescription of processes, the *C* theory appears to trivialise time reversal invariance since it is not immediately clear what sense can be made of time reversal *non*-invariance on the *C* theory. The problem is statable as a simple argument:

P1. In order for a theory to be time reversal non-invariant, a model of the theory must transform under time reversal to a non-model of theory. [assumption]

P2. On the *C* theory of time, time reversal is just a redescription of a single possible world. [definition]

P3. A possible world cannot be deemed by some theory to be 'physically possible' relative to one description and 'not physically possible' relative to an equivalent description. [assumption]

C. Hence, on the *C* theory, no theory can be time reversal non-invariant. [P1, P2, P3]

It might be thought that this implies that the difference between reversible and irreversible theories is not statable in *C*-theoretic terms, which would be a major weakness for the *C* theory—we evidently do have a clear grasp on the differences between reversibility and irreversibility, as well as other kinds of probabilistic time asymmetries, so the *C* theory had better possess the resources to

[21] Indeed, Earman points to this distinction:

> [T]he Reichenbach–Gold position [i.e. the *C* theory] cannot be based solely on time reversal invariance, but must rely on specialized assumptions about the nature of time reversal invariance. These assumptions have never been explicitly stated, much less justified. (Earman, 1974, p. 24)

I should add that these specialised assumptions concern *time reversal* and not *invariance* under time reversal *per se*.

account for this. However, the argument does not pose such a problem for the *C* theory: irreversible processes are straightforwardly describable in *C*-terms. As I show, though valid, the main problem on which the argument picks up is that time reversal as applied to entire models of a theory is insufficient to determine whether a theory describes irreversible processes, which I shall argue is a problem for both the *B* and *C* theories. Let's first establish that irreversibility is statable in *C*-terms.

Consider a time-asymmetric law describing the behaviour of some variable x:

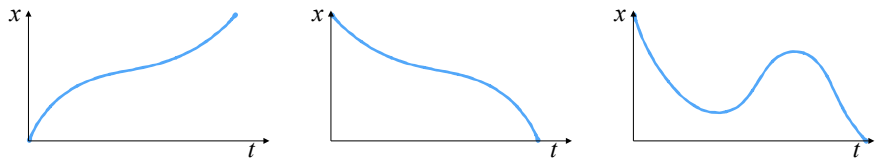
L. The value of x increases, and never decreases, with time.

This describes an ideal irreversible process: the increase in value of x.[22] There are two notable features of L relevant to time reversal and the *B* and *C* theories. First, L describes an irreversible process. Second, L is stated in time-directed (*B*-theoretic) terms: if x increases relative to one temporal direction, it *decreases* relative to the opposite temporal direction, and thus L is not stated in a time-direction-neutral way. For the *C* theorist, L is equivalent to L*:

L*. The value of x decreases, and never increases, with time.

If we take a model m that satisfies L, then its TR-twin m^* satisfies L*. The *C* theorist takes TR-twins m and m^* to represent a single possible world, and so to privilege neither L nor L*. However, this does not mean that the *C* theory is unable to accomodate the irreversibility described by L and L*. There is a key sense of irreversibility that is independent of time direction and so statable in *C*-terms: the x-process described by L (and by L*) is *monotonic*. For some type of process to satisfy either L or L* it must be monotonic, such that relative to a choice of positive time it either: (a) only increases; or (b) only decreases. The x-process is monotonic regardless of whether we take it to be an 'x-increasing' or 'x-decreasing' process. Figure 1 depicts three models to illustrate this: fig. 1a and 1b depict the TR-twins m and m^*, and fig. 1c depicts a model, n, in which the value of x changes non-monotonically. On the *C* theory, although m and m^* represent a single possible world, they are structurally distinct from n. Importantly, fig. 1a and 1b depict a monotonic gradient regardless of the designation of a direction of time. As such we can offer a *C*-theoretic version of L:

[22] Though I use an irreversibility as an illustrative example, the following line of reasoning equally well applies to probabilistic time asymmetries that are weaker than strict irreversibility.



(a) m (b) m^* (c) n

Figure 1: Models of monotonic variation (a and b), and non-monotonic variation (c), of a variable, x. (a) and (b) are TR-twins, m and m^*: (a) m represents the monotonic increase of x; (b) m^* represents monotonic decrease of x. (c) Model n represents a non-monotonic x process.

L^C. The value of x changes monotonically in time.

Such a law requires all x-processes to be coordinated such that relative to a choice of positive time they are either all x-increasing or all x-decreasing. This key sense of irreversibility is thus statable in *C*-terms.

This point may be generalised to non-idealised cases of putatively irreversible processes, such as the statistical time asymmetry of thermodynamics, and the time asymmetry of dynamical collapse theories such as the Ghirardi–Rimini–Weber theory (GRW), and also of probabilistic time asymmetries in particle physics, such as the decays of K^0 and B^0 mesons.[23] Regarding thermodynamics, there is a clear conventional element in taking entropy to increase; for all we know, it could be that time really 'goes' from our future to our past, and hence it be a law that entropy tends to *de*crease, and never increase, contrary to our beliefs. What is important in accounting for the phenomena is not whether entropy 'really' increases or decreases, but rather that, *once we've fixed our convention* about the positive direction of time, entropy either 'increases and does not decrease' or 'decreases and does not increase'.[24] Both time-direction-dependent descriptions pick up on the time-direction-independent monotonicity of entropy—entropy doesn't fluctuate in both temporal directions.[25] The *C* theory captures this sense of irreversibility.

[23] The experimental violation of the combination of charge and parity symmetry (CP symmetry) in particle is well-documented. For a discussion of CP violation in K^0 meson decay, see Sachs (1987, chs. 8–9); for B^0 mesons, see Abe et al. (2001).

[24] This point is made at length by Price (1996b), particularly chapters 2 and 7.

[25] With respect to GRW, the key content of the irreversibility of collapses is that the set of collapses are co-oriented with respect to time such that in each GRW model there is, relative to a choice of time direction, either: (a) only collapses; or (b) only 'anti-collapses'.

The *B* theory additionally legitimises the question of whether entropy might 'really' increase-and-not-decrease or decrease-and-not-increase, but this is a separate issue that is not clearly epistemically accessible, nor for that reason practically indispensable to the study of time asymmetry.[26] As such, not only is law-like irreversibility and time asymmetry statable in *C*-terms, but is also better understood in such terms.

With this in mind, let's return to the argument on p. 11. What is problematic regarding time reversal and irreversibility is that if some model satisfies L^c then so does its TR-twin, regardless of whether one takes the TR-twins to represent different possible worlds, and so this is a problem for both the *B* and *C* theories. P1 states a requirement for a theory to be time reversal non-invariant that appears to be satisfied only by the *B* theory and not by the *C* theory, since only the *B* theory treats time reversal as an active transformation and so TR-twins as describing different possible worlds. On the *B* theory—but not the *C* theory—we can understand m and m^* as representing distinct possible worlds, and so the *B* theory allows in principle for a theory to contain m as a model without also containing m^* as a model. However, failure of time reversal invariance in this sense would be quite odd. First, the choice of which model m or m^* is used to represent some process is a matter of convention,[27] and so a theory's inclusion of one and not the other in its space of models would also be a matter of convention. Moreover, second, if a theory includes only one of a pair of TR-twins, it does not follow that the theory contains any lawlike irreversible or probabilistically time-asymmetric processes. For lawlike time asymmetry, a theory would have to satisfy a stronger condition. For example, a theory would contain a lawlike irreversibility in the case that for some variable x, the theory includes models of monotonic x-processes—such as m and m^*—, and excludes all models of non-monotonic x-processes—such as n.

It is because of this second point that the argument on p. 11 is misleading. The argument establishes that the *B* theory allows for a theory to be time reversal non-invariant in a way in which the *C* theory does not, but this is only in the case that time reversal is understood as an operation upon an entire model of a theory. However, time reversal applied to an entire model cannot transform monotonic models such as m and m^* to non-monotonic models such as n, and

[26] Even supposing there is a privileged direction of time along which processes 'really' occur, we evidently do not need knowledge of this to collectively prefer to say that entropy 'increases' rather than 'decreases'.

[27] In other words, we could in principle have preferred to describe processes in our world from future-to-past rather than from past-to-future without getting anything 'wrong'.

so non-invariance under such an operation is insufficient as a test for law-like irreversibility. This follows from our pragmatic constraint that time reversal functions such that TR-twins represent what a process 'looks like' relative to the opposite time directions—it would be unreasonable given this for time reversal to fail to preserve monotonic and non-monotonic behaviour. As such, P2 in particular requires clarification; it is important to stress the context of the *C* theorist's claim that time reversal amounts to a redescription of processes. If we take a model of an 'expanding' gas, such that its TR-twin is a 'contracting' gas, the *C* theory entails that these are equivalent descriptions of a gas occupying greater volume at one temporal end than the other. However, were we to embed this model in a wider environment containing other gases displaying matching time-asymmetric behaviour, things would be different. In this case, switching one model for its TR-twin and holding the orientation of the other gases fixed would result in a physical change to the total system (e.g. so that there were now a gas 'contracting' relative to the time direction in which the other gases were 'expanding')—it would constitute a change to the *C* series, not only to the *B* series, of the total system, since it would amount to a difference in the temporal-betweenness ordering of events, and not only the earlier-than ordering. This sense of *relative* time reversal corresponds to an active change even on the *C* theory.

This gives us two different kinds of time reversal. First, a relative time reversal is an active transformation on both the *B* and *C* theories, since it changes the temporal betweenness relations, and hence constitutes a change regardless of stipulation of time direction.[28] Second, an absolute time reversal—applied to an entire model, or to a total system (e.g. the entire world)—is a passive transformation on the *C* theory and an active transformation on the *B* theory; only on the *B* theory does a world identical to ours save for the direction of time constitute a different possible state of affairs. The argument on p. 11 establishes that only the *B* theory allows for a theory non-invariant with respect to absolute time reversal. However, non-invariance under absolute time reversal is neither necessary nor sufficient for a theory to contain a lawlike time asymmetry or irreversibility, and as such the *B* and *C* theories equally well account for the existence of time asymmetries and irreversible processes.

[28] Note however that on the *B* theory relative time reversal (change of the *C* series) can in principle be carried out in two different ways: first, by holding the lab's time orientation fixed and time reversing the experimental system; second, by holding the experimental system's time orientation fixed and time reversing the lab.

2.3 Answers

To recap, both the *B* and *C* theories can support physical laws that describe irreversible or probabilistically time-asymmetric processes. The central point at issue between the two theories is whether the application of a set of time reversal transformations to a model of some theory takes us to a model that describes a different possible world. Time reversing an entire model amounts to a redescription of a single possible world according to the *C* theory, but amounts to a description of a second, distinct possible world according to the *B* theory. On the *C* theory, time reversal is a purely passive, coordinative transformation, meaning TR-twins differ only in terms of notation—they represent a single possible world, and hence notation that varies under time reversal (such as the direction of velocities) should not be taken to represent a property of the target system. On the *B* theory, time reversal involves altering fundamental temporal relations and hence takes us from one logically possible world to a distinct logically possible world.[29] I have argued that the extra structure postulated by the *B* theory is not required to account for temporally asymmetric phenomena.

With regard to the relationship between causation and time reversal, we can now clarify the problem for causation posed by CTR. According to CTR, TR-twins represent different sets of causal relations. On the *C* theory, since TR-twins represent a single possible world, the only way for TR-twins to agree on causes and effects, given CTR, is for there to be no causes or effects, and hence this motivates causal eliminativism. The *B* theory, conversely, contains the logical space for TR-twins to disagree over causal direction facts insofar at they represent different possible processes in different possible worlds, and as such the cause–effect asymmetry can be preserved in each possible world. However, in the next section we see that the combination of the *B* theory and CTR leads to major problems in paradigm cases of causal processes, and as such I propose that ¬CTR should be preferred.

3 Does time reversal reverse causal relations?

We may now ask whether time reversal inverts causal relations. This section examines the relative appeal CTR and ¬CTR in the context of (1) a time-symmetric process and (2) a time-asymmetric process.

[29] Whether or not these *logically* possible worlds are deemed *physically* possible depends upon whether the relevant physical theory is time reversal invariant.

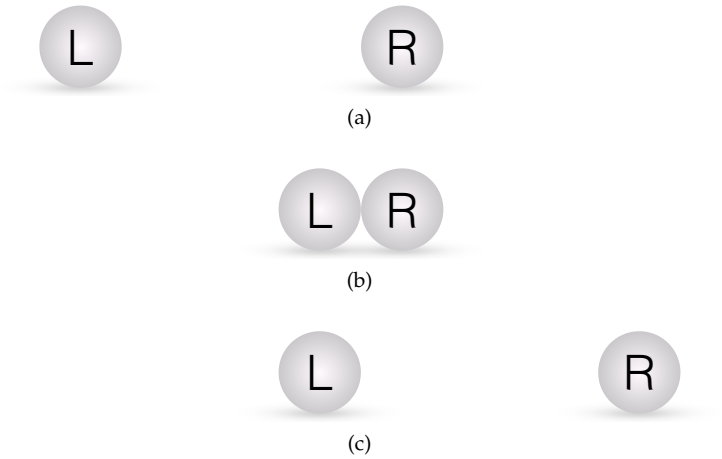


Figure 2: The time symmetric process of a collision of two idealised snooker balls of equal mass on a frictionless plane.

3.1 Causation, billiards and snooker

3.1.1 Causation and time-symmetric processes

Figure 2 depicts the time-symmetric process of a collision of two idealised billiard balls of equal mass on a frictionless plane. In 2a, ball L has non-zero momentum and ball R is at rest. At 2b there is a perfectly elastic collision, upon which the total momentum of one ball is transferred to the other. Figure 2c depicts ball L at rest and ball R with non-zero momentum. From 2a to 2c, ball L's movement appears to cause ball R's movement, and from 2c to 2a, ball R's movement appears to cause ball L's movement. Assuming CTR, if the 2a–2c account represents a causal process in which ball L's momentum causes ball R to move, then 2c–2a represents the distinct causal process in which ball R's momentum causes ball L to move. To fill in these distinct accounts, we can imagine a right-pointing arrow from L to R in 2a on the 2a–2c process, and a left-pointing arrow from R to L in 2c in the 2c–2a process. Assuming ¬CTR, both 2a–2c and 2c–2a represent the same causal process.

As we've seen, CTR and ¬CTR relate differently to the *B* and *C* theories. On the *C* theory, since time reversal amounts to a redescription of a single possible process, CTR is untenable since it requires 2a–2c and 2c–2a to represent distinct possible processes. Hence the *C* theory requires ¬CTR, and CTR requires the *B* theory. The combination of the *C* theory and ¬CTR applied to the billiards

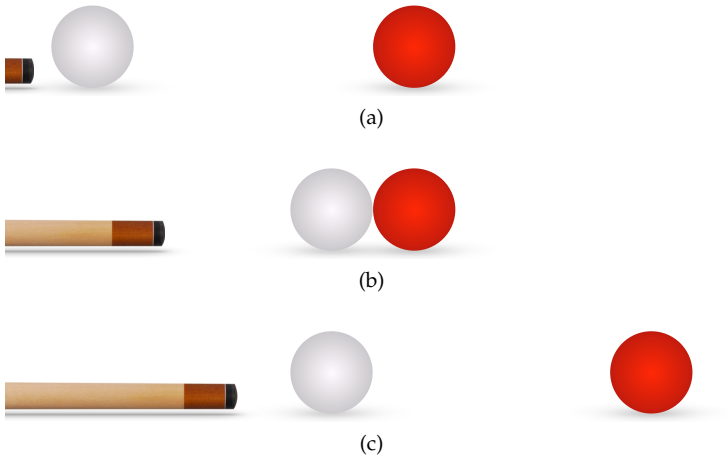


Figure 3: The time asymmetric collision of two realistic snooker balls of equal mass on a frictional snooker table.

example suggests that if there is a causal process described here, the direction of causation is at best ambiguous. (We examine this issue in sec. 3.2.) On the *B* theory, since 2a–2c and 2c–2a represent different possible processes, it is natural to read them as distinct causal processes. The combination of the *B* theory and CTR fits with microphysical accounts of causation, such as dispositional accounts, in which causation is understood as some kind of unidirectional influence from one object to another. On such an account, 2a–2c and 2c–2a refer to fundamentally different processes. In the 2a–2c process the initial non-zero left-to-right momentum of ball L is a cause of the collision. In the 2c–2a process the 'initial' non-zero right-to-left momentum of ball R is a cause of the collision.[30]

3.1.2 Causation and time-asymmetric processes

Figure 3 depicts the time-asymmetric collision of two realistic snooker balls of equal mass on a frictional snooker table. In this case, an element of agential control is introduced: there is a snooker cue that interacts with the white cue ball. Furthermore, the presence of a non-conservative force—friction—brings in an important explanatory asymmetry between the 3a–3c and 3c–3a accounts, and it is more convenient to describe the process in time-directed terms, unlike in the time-symmetric case.

In the conventional 3a–3c description, the cue strikes the cue ball, setting it

[30] It is also tenable for the *B* theorist to adopt ¬CTR, as is discussed in sec. 4.2.

in motion, and the cue ball then collides with the red ball, transferring most of its momentum to the red ball. The red ball then loses momentum due to the frictional force of the baize on the snooker table until it is at rest, as depicted in 3c. The 3a–3c description contains a number of causal terms, implying the following: the cue movement causes the cue ball's movement; the cue ball's movement causes the red ball's movement; the baize causes the red ball to lose momentum.

In the unconventional 3c–3a description, an anomalous series of causal processes is implied. Firstly, heat in the baize together with incoming air molecules conspire to set the red ball in motion. Secondly, the red ball's motion in synchrony with inverse, concentrating soundwaves jointly impart a gain in momentum in the collision of the red ball into the cue ball. Finally, the cue ball's momentum is absorbed in a collision with the cue. As a candidate causal process, 3c–3a is highly unsatisfactory. Two issues in particular stand out: (1) there are several points that imply a violation of the Causal Markov Condition (CMC); (2) the snooker player apparently loses her agential control over the balls' motion. These imply both a causal and explanatory asymmetry between the two available time-directed descriptions 3a–3c and 3c–3a, which, as I next argue, motivates ¬CTR.

3.2 The epistemology of causal direction

The 3c–3a description, understood as a causal process, implies the existence of causally independent variables that nonetheless exhibit coordinated behaviour and hence are not statistically independent, in violation of the CMC.[31] In simpler terms, there are a number of coincidences that can't be explained away with reference to some common interactions in the causal past. As such, there is good reason to think that this does not represent a genuine causal process: it does not meet a standard criterion widely taken to be characteristic of causal relations, and central to the explanatory asymmetry of causes and effects.[32]

[31] The statistical dependence here is merely implicit. Given that the example depicts only a single (though abstract) run of the process, there is merely an apparent coincidence in that the initial conditions are highly improbable — they appear fine-tuned to entail coordinated behaviour. On multiple runs of this exact scenario, the statistics produced would provide a straightforward violation of the CMC, which holds that causally independent variables (relative to their causal pasts) are statistically independent — cf. Hausman and Woodward (1999).

[32] This point can be quite easily restated in terms of Lewis's (1979) counterfactual theory of causation: the coincidences in the 3c–3a process entail that in such a case the past 'overdetermines' the future, and as such there is counterfactual dependence of earlier events upon later events and not *vice versa* on Lewis's possible worlds semantics.

Calling such a process 'causal' is to insist on using the term quite outside its standard linguistic context and is thus heuristically unhelpful. After all, in order for the concept of causation to be useful in philosophical discourse, there ought to be reasonable restrictions on its domain of application so to exclude processes that violate standard causal criteria such as CMC and its variants.[33] For this reason, it is useful to defer to the patterns of conditional dependencies and independencies of variables to ascertain causal direction, as is characteristic of causal modelling.[34]

Furthermore, the introduction of agential control brings in a pragmatic constraint on causal inference: it is natural to stipulate that the snooker player has causal control over the cue and of the cue ball, and not *vice versa*. One can entertain a 3c–3a causal process whereby the cue's movements are (at least in part) caused by the motion of the cue ball, but it detaches various causal intuitions we have about snooker players from the causal relations described in the account. Reichenbach considers a similar problem regarding, in his case, tennis players and time-reversed 'causal' processes:

> It would be a strange experience indeed to see [tennis] players run backward. Such a motion, although compatible with the laws of mechanics, is unusual because we are safer if our steps are controlled by our eyes. (Reichenbach, 1956, p. 47)

This element of control is important in that it can be appealed to in order to privilege one of the two possible causal stories given relative to the opposite directions of time. Regardless of any underlying time symmetry, and regardless of any freedom to describe some process relative to either time direction, it is desirable to hold that we are not mistaken in such control judgements. This is because the appeal to control plays an explanatory role: it is reasonable to take the snooker player's actions to explain the subsequent motion of the snooker balls and not *vice versa*.

The notion of control and manipulation are central to agency and interventionist theories of causation, such as those of Menzies and Price (1993), Pearl (2000) and Woodward (2003). These provide a deflationist epistemology of the direction of causation, whereby the direction of causation is determined by the kind of patterns of correlations to which causal discovery algorithms are sensitive. In the case of fig. 3, we can appeal to the CMC, or more prosaically ap-

[33] E.g. common-cause principles and screening-off conditions—cf. Arntzenius (2010).
[34] cf. Pearl (2000); Spirtes et al. (2001); Woodward (2003).

peal to beliefs about the snooker player's agential control, to ground a direction of causation.[35] A deflationist account of causal direction holds that there is a direction of causation only in the presence of the right kind of probabilistic asymmetries (e.g. irreversible processes, time-asymmetric screening-off conditions, etc.). Although the deflationist approach is applicable to our agential snooker case, it leaves open the status of causation in our idealised billiards case, in which there are insufficient asymmetries to ground a direction of causation. One option is that there just is no direction of causation intrinsic to such time-symmetric systems, but if one can refer to a wider system containing (for example) irreversible processes then this can be used to define a direction of causation in the time-symmetric system. Such problem cases need not worry us in practice, since in general we do have sufficient asymmetric processes (e.g. ourselves) to which to refer.[36] In the idealised case of a world consisting solely of our idealised billiards example, the deflationist may hold that there is no fact about causal direction.[37] Such an attitude towards idealised time symmetric systems does not entail eliminativism nor scepticism about the direction of causation in worlds containing sufficient time asymmetries to determine a direction of causation. As such, the compatibility of such worlds with physical theories that are empirically adequate with respect to our world does not motivate eliminativism about causal direction with respect to our world.

Whereas ¬CTR aligns with a deflationist account of causal direction, CTR aligns with a hyperrealist account of causal direction, whereby there is a causal direction that both outruns and is independent of the physical facts.[38] This is because in order for time reversal to invert the direction of causation, the direction of causation must be independent of time-independent causal algorithms that ground the deflationist account of causal direction.[39] To point to an ex-

[35]Stipulations about agency and control play a key constitutive role in causal modelling. In general, multiple causal models will be compatible with the statistical data concerning relationships between variables of a system, and designating certain variables as 'exogenous' (i.e. 'free' variables that are not effects of other variables in the system) narrows down the set of viable causal models for the system.

[36]See Farr (2016) for a discussion of this issue in the context of the debate between John Norton (2009) and Mathias Frisch (2009, 2014) about causal reasoning in time symmetric systems.

[37]The *C* theorist could instead commit to a symmetric notion of 'causal betweenness', which provides an ordering that is invariant across TR-twins. This route appears to be taken by Reichenbach (1956, p. 191).

[38]See Price and Weslake (2010) for a critique of hyperrealist accounts of the direction of causation.

[39]The patterns of statistical (in)dependence with which causal discovery algorithms are concerned are themselves neutral with respect to the direction of time. For instance, retrocausality—whereby a pair of cause/effect events are such that the cause event is later relative to clock time (e.g. of a lab) than the effect event— is conceptually possible relative to such algorithms.

ample of hyperrealism, Maudlin (2007) takes the direction of causation to be determined by the 'passage of time', which he regards as "an ontological primitive [that] accounts for the basic distinction between what is to the future of an event and what is to its past" (*ibid.*, p. 172). As such, the direction of causation is independent of any particular probabilistically time-asymmetric processes in the world: "[causal] production [is] built on the foundational temporal asymmetry that would obtain even if the world were always in thermal equilibrium (even then, later states would arise out of earlier ones)" (*ibid.*, p. 177). A hyperrealist account might seem preferable in the idealised billiards case, since the deflationist approach is silent about causal direction. However, the hyperrealist is still faced with the epistemic problem faced by the deflationist: there is no clear causal direction to be derived from the physical facts.[40] Rather, the hyperrealist approach here is to stipulate a preferred causal arrow to artificially break the symmetry. While this may be innocuous in the billiards case, it creates a significant problem in cases like the agential snooker example where we have objective physical grounds for determining a preferred arrow of causation. Since the hyperrealist approach is by its nature insensitive to the kinds of factors that inform causal judgements, it gives up the explanatory benefits of the deflationist approach. Taking the direction of causation to be an ontological primitive licenses worries about whether the snooker player's action 'really' causes the movement of the snooker balls or vice versa, which is not a legitimate worry on the deflationist approach. In the kinds of cases where we naturally make unambiguous causal judgments, such as the snooker case, the deflationist epistemology of causation of ¬CTR is preferable to the hyperrealism of CTR. As such, causal relations should not be taken to reverse under time reversal.

3.3 Answers

If we are to consider archetypal causal processes, namely those that satisfy standard algorithms for causal discovery, then we ought to hold that causal relations do not invert under time reversal, and so prefer ¬CTR to CTR. Though it may be intuitively plausible for causal relations to reverse under time reversal, such a view is reasonable only with respect to suitably time-symmetric cases—like the idealised billiards case of fig. 2—where there is no clearly preferred direction of causation. I have argued that in such cases it is better to be neutral with

[40] As Price and Weslake (2010) argue, hyperrealism about causation requires a denial of physicalism.

respect to causal direction rather than to adopt a hyperrealist account of causal direction.

¬CTR fits naturally with a *C* theory of time. Combining the two, one may consider the two temporally-opposed descriptions of the agential snooker example as equivalent descriptions of a single possible causal process. The 3c–3a description, though unconventional in its form, may to be taken to represent the same causal relations that are naturally read from the 3a–3c description. The asymmetry between the two descriptions is not due to any important link between causation and time, but rather due to time-independent factors that inform causal judgements. The issues of agency and the CMC lead to the same judgements about causal direction regardless of what one takes to be the underlying direction of time. This entails that any underlying time-reversal invariance of the microphysical description is beside the point: one may hold that there is a clear causal direction, a natural criterion for distinguishing between causes and effects in the example, which is invariant under time reversal.

4 Is time reversal symmetry compatible with causation?

We are now in a position to evaluate the central question of the paper: is time reversal symmetry compatible with causation? In the previous sections, we considered the following questions:

- Do TR-twins represent distinct possible worlds?
- Does time reversal invert causal relations?

These present four options, as listed in table 1. It follows from our considerations that options 2–4 give us compatibilism about causation and time reversal symmetry, and that of these, Option 3 (¬CTR and the *C* theory) is the preferred option. Before reviewing the compatibilist options, we can first look at the incompatibilism of Option 1.

4.1 Incompatibilism

Option 1: CTR + *C* theory = Incompatibilism

I have suggested that, assuming the *C* theory, if there are directed causal relations between events then these cannot be flipped under time reversal. I have

Table 1: Table of options.

	C Theory	*B* Theory
CTR	Option 1	Option 2
¬CTR	Option 3	Option 4

taken this to show that the *C* theory requires a non-causal understanding of time reversal. Interestingly, Gold (1966) appears to go in the opposite direction and take his passive (*C*-theoretic) interpretation of time reversal to entail a Russellian causal eliminativism, holding that "[t]he idea of a cause and effect relationship now becomes meaningless" (Gold, 1966, p. 327). Gold's contention is based on a *causal* interpretation of time reversal:

> You may see relationships within [a time-direction-neutral description] which are of the kind that in the conventional description one would be called the cause and the other the effect. In the description with the opposite sense of time you would *just have to reverse these roles*. (Gold, 1966, pp. 327–8; my emphasis)

Given that the *C* theory lacks the structure to commit to two such worlds with distinct causal relations, applying CTR does indeed entail eliminativism: the only way for TR-twins to agree on causes and effects, assuming that these are flipped by time reversal, is for there to just be no causes or effects. Conversely, if a *C* theorist *does* want to commit to directed causal relations, then these must be fixed by properties of the *C*-theoretic model expressible in time-direction-neutral terms, and thus left invariant under time reversal. Seen in this way, the *C* theorist is committed to no causal relations being flipped by time reversal and thus to ¬CTR, *contra* Gold.

The central problem is that the following three claims form an inconsistent triad:

1. There are directed causal relations between events.
2. Time reversal reverses causal relations. [CTR]
3. TR-twins describe the same possible world. [C theory]

Though each statement is independently plausible, the three jointly entail a contradiction. However, as we have seen, we may reject any one of these claims and

avoid inconsistency. As should be clear, I take claim 2 to be the one to reject. What is most important though is that either 2 or 3 may be rejected so to save 1. The mutual incompatibility does not mark out 1 as being the problematic claim.

4.2 Compatibilism

Option 2: CTR + *B* theory = Compatibilism

Option 2 avoids incompatibility by rejecting claim 3 of the triad (the *C* theory). The *B* theory holds that TR-twins describe distinct possible worlds, and this provides the logical space for there to exist directed causal relations that are flipped by time reversal without engendering a contradiction: in each *B*-theoretic world, the asymmetry of cause and effect is preserved. In place of the direct incompatibility of Option 1, Option 2 gives us practical and epistemological problems concerning directed causal relations in the kinds of cases in which we routinely make unambiguous causal judgements, such as in the snooker example (fig. 3). In allowing the sequences of fig. 3a–3c and fig. 3c–3a to represent distinct causal processes, Option 2: (a) leads to a problem of underdetermination, since both 'causal' processes are consistent with the same sets of data; and more importantly (b) fails to account for why 3a–3c and 3c–3a are asymmetric with respect to explanation in that only the former satisfies standard algorithms for causal discovery. These problems are unique to this approach. It is only by committing to CTR that the causal realist can entertain the possibility of processes whose causal direction is the opposite to that given by causal discovery algorithms.

Option 3: ¬CTR + *C* theory = Compatibilism

I have argued that Option 3 is the preferred account: by holding that TR-twins represent the same possible world (*C* theory) and that cause and effect is invariant under time reversal (¬CTR), one can hold the time reversal invariance of a theory to pose no conceptual or epistemological problem for the direction of causation. We've seen that Option 3 has several key benefits. First, the key sense of lawlike time-asymmetry that is satisfied by irreversible or probabilistically time-asymmetric processes is captured by the *C* theory. Second, combining the *C* theory with ¬CTR allows for a deflationist epistemology of causal direction that (i) preserves causal direction judgements as determined by standard causal algorithms, and (ii) dissolves scepticism as to whether the direction of causation matches our standard causal direction judgements.

Table 2: Is time reversal symmetry compatible with causation?

	C theory	*B* theory
CTR	✗	✓
¬CTR	✓	✓

Option 4: ¬CTR + *B* theory = Compatibilism

The final option, which I have not discussed up to this point, is to combine the non-causal account of time reversal with the *B* theory. There are in principle a couple of ways to do this: (1) defend a primitivist account of the direction of causation and stipulate that this should not be inverted by time reversal; (2) defend the same epistemology of causal direction as that of the *C* theorist, but additionally hold that TR-twins describe distinct worlds with different time-direction facts. This second option preserves the epistemic advantages of my preferred option—Option 3—, but additionally allows for two worlds to differ solely in terms of 'earlier than' relations. In this sense, the *B* theorist can avoid the epistemological problems faced in Option 2. However, this then requires that the direction of time is wholly independent of the direction of causation. This kind of realism about the direction of time may have independent motivations and benefits that are outside the scope of this paper. However, in terms of the cases we've considered, I take the *C* theory to be the natural metaphysics of time for a non-causal interpretation of time reversal.

4.3 Answers

Time reversal symmetry is compatible with the existence of directed causal relations. However, realism about causal direction comes with restrictions as shown in our inconsistent triad: either CTR or the *C* theory must be rejected, as summarised in table 2. Moreover, I have argued that the most reasonable resolution of the triad is to reject CTR: time reversal should not be understood as inverting causal relations.

Crucially, the compatibility of time reversal symmetry and causation depends upon the interpretation of time reversal itself and is independent of whether any particular physical theory is invariant under time reversal, contrary to standard presentations of the Directionality Argument. In our incompatibilist option—Option 1—the incompatibility is due to the combination

of the CTR and the *C* theory. For the compatibilist options, compatibility is due *either* to holding that time reversal does not invert causal relations (¬CTR) *or* to holding that TR-twins represent distinct possible worlds (*B* theory). Each option is consistent with both time reversal invariant and time reversal non-invariant theories.

5 Outlook

Causation and time reversal invariance are not straightforwardly incompatible. Rather, the relationship between the two depends upon one's interpretation of time reversal. I've shown that there are several compatibilist options available to the causal realist. Moreover, I have argued in favour of both the *C* theory and ¬CTR: time reversal should be understood as a passive transformation that re-describes a single possible world, and so time reversal does not invert causal relations. This entails a suggestion about how to think of properties of instantaneous states that are acted upon by time reversal operations: such properties (e.g. velocity, momentum, etc.) are either (a) not causal, or (b) not genuine properties of instantaneous states. That is to say, we cannot take a naive view of velocities or momenta as telling us something about the direction of causal propagation or information flow. We can *either* take velocities to be non-causal in nature, so that velocities do not amount to something like causal dispositions—they just point one way or the other without contributing to the causal structure of a system—*or* we can take the direction of the velocity of a particle to be fixed by its position in a wider causal environment in which causal relations can be determined relative to causal discovery algorithms. This suggests a certain contextuality of such quantities—x has some velocity only relative to causal model Y.[41]

If we take CTR and the *C* theory to both be appealing, which is reasonable, then we might think that causation is eliminated. However, this tacitly presupposes that causal facts are to be found in the microdynamics in the first place. I think this is to start off on the wrong foot. We should think that insofar as

[41] Price (1996a) suggests this kind of case as a problem for reducing causal direction to the fork asymmetry of causal models in microphysics: the direction of a causal process will ultimately be determined by which variable one includes in one's model. This is suggestive of an arbitrariness of causal direction as determined by causal models. In particular, an open issue for this approach is what to make of *postselected* causal models, whereby the data is chosen in such a way to reveal patterns of correlations suggesting the opposite causal direction to that which we take to hold in the world. It is interesting as to whether we can reject the significance of such apparently causal relations on grounds of being artificial or unnatural. This is an issue for another paper.

we do make causal direction judgements and wish to ascribe them to physical systems, these judgements derive from higher-level statistical observations and agential presuppositions that are themselves neutral regarding any micro-dynamical arrows of time or causation. As such, the time reversal symmetry of the underlying dynamics need not require us to doubt whether there really are directed causal relations. This is welcome, since it is quite reasonable to be ambivalent about whether fundamental physics is time reversal invariant. After all, the world of our experience accords to time reversal non-invariant laws (e.g. thermodynamics), which are underpinned by the time reversal invariant laws of classical physics, which themselves are an approximation of quantum mechanics, which is on many popular formulations time reversal non-invariant. It is desirable to avoid such worries when considering the status of causation.

References

Abe, K., R. Abe, I. Adachi, B. S. Ahn, H. Aihara, M. Akatsu, G. Alimonti, K. Asai, M. Asai, Y. Asano, et al. (2001). Observation of large cp violation in the neutral b meson system. *Physical Review Letters 87*(9), 091802.

Albert, D. Z. (2000). *Time and Chance*. Massachusetts: Harvard University Press.

Arntzenius, F. (2010). Reichenbach's common cause principle. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010 ed.).

Arntzenius, F. and H. Greaves (2009). Time reversal in classical electromagnetism. *British Journal for the Philosophy of Science 60*(3), 557.

Black, M. (1959). The "direction" of time. *Analysis 19*(3), 54–63.

Black, M. (1962). *Models and Metaphors: Studies in Language and Philosophy*. Ithaca: Cornell University Press.

Callender, C. (2000). Is time 'handed' in a quantum world? *Proceedings of the Aristotelian Society 100*(1), 247–269.

Earman, J. (1974). An attempt to add a little direction to "the problem of the direction of time". *Philosophy of Science 41*(1), 15–47.

Farr, M. (2016). Mathias Frisch: Causal Reasoning in Physics. *The British Journal for the Philosophy of Science*.

Farr, M. (MS). The C theory of time. unpublished manuscript.

Farr, M. and A. Reutlinger (2013). A Relic of a Bygone Age? Causation, Time Symmetry and the Directionality Argument. *Erkenntnis 78*(2), 215–235.

Field, H. (2003). Causation in a physical world. In M. Loux and D. Zimmerman (Eds.), *Oxford Handbook of Metaphysics*, pp. 435–60. Oxford: Oxford University Press.

Frisch, M. (2009). Causality and dispersion: A reply to John Norton. *The British Journal for the Philosophy of Science 60*(3), 487–495.

Frisch, M. (2012). No place for causes? Causal skepticism in physics. *European Journal for Philosophy of Science 2*(3), 313–336. 10.1007/s13194-011-0044-4.

Frisch, M. (2014). *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Gold, T. (1966). Cosmic processes and the nature of time. In R. G. Colodny (Ed.), *Mind and Cosmos*, pp. 311–329. Pittsburgh: University of Pittsburgh Press.

Hausman, D. and J. Woodward (1999). Independence, invariance and the causal markov condition. *The British Journal for the Philosophy of Science 50*(4), 521–583.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs 13*(4), 455–476.

Maudlin, T. (2007). *The Metaphysics Within Physics*. Oxford: Oxford University Press.

McTaggart, J. M. E. (1908). The unreality of time. *Mind 17*(68), 457–474.

McTaggart, J. M. E. (1927). *The Nature of Existence*, Volume II. Cambridge: Cambridge University Press.

Menzies, P. and H. Price (1993). Causation as a secondary quality. *The British Journal for the Philosophy of Science 44*(2), 187–203.

Ney, A. (2009). Physical causation and difference-making. *British Journal for the Philosophy of Science 60*(4), 737.

Norton, J. D. (2009). Is there an independent principle of causality in physics? *The British Journal for the Philosophy of Science 60*(3), 475–486.

Oreshkov, O. and N. J. Cerf (2015, 07). Operational formulation of time reversal in quantum theory. *Nature Physics advance online publication*, –.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Price, H. (1996a). Backward causation and the direction of causal processes: Reply to Dowe. *Mind 105*(419), pp. 467–474.

Price, H. (1996b). *Time's Arrow and Archimedes' Point: New directions for the physics of time*. Oxford: Oxford University Press.

Price, H. and B. Weslake (2010). The time-asymmetry of causation. In H. Beebee, C. Hitchcock, and P. Menzies (Eds.), *The Oxford Handbook of Causation*, pp. 414–443. Oxford: Oxford University Press.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

Russell, B. (1912–1913). On the notion of cause. *Proceedings of the Aristotelian Society 13*, pp. 1–26.

Sachs, R. G. (1987). *The Physics of Time Reversal*. Chicago: University of Chicago Press.

Spirtes, P., C. Glymour, and R. Scheines (2001). *Causation, prediction, and search*. The MIT Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.