Counterfactuals, Irreversible Laws, and the Direction of Time

> There is another very subtle law of physics that may be even more fundamental than energy conservation. It's sometimes called reversibility, but let's just call it *information conservation*. Information conservation implies that if you know the present with perfect precision, you can predict the future for all time. But that's only half of it. It also says that if you know the present, you can be absolutely sure of the past. It goes in both directions.
>
> Leonard Susskind [1]

The doctrine Susskind calls 'Information Conservation', philosophers call 'Determinism'.

> Determinism
>
> For any physically possible world: the laws of nature together with a description of the total state of the world at any time, entail a description of the total state of that world at every time.

Here Susskind describes information conservation as a "law" more fundamental than energy conservation; elsewhere he has said it is the most fundamental law of classical mechanics.[2] But unlike 'energy' and 'heat, 'information' does not have a direct physical metric. Information Conservation is not a claim about some physical quantity but a second order claim about what the laws of nature must be like. It is better viewed as what Kant would have called a "Regulative Principle" rather than a law. It expresses a completeness condition which we expect any total theory of the world to satisfy: we require a theory of the world to allow us, in principle, to predict and retrodict the state of the world at every moment, given only a description of its state at any particular moment. If it cannot, then we will regard the theory as wrong, or at least incomplete.[3]

Our most comprehensive theories are deterministic or at least aspire to be. Classical physical theories from Newton to Einstein are deterministic (though by the time we get to General

---

[1] The Black Hole War. p. 86.

[2] Susskind, 2007.

[3] We might be wrong about this. It is a contingent matter whether reality will conform to this expectation. But it is not clear that we could ever *know* that we were wrong. It is difficult to imagine finding empirical evidence that would prove that there was no better deterministic theory to be had.

Nb. This essay reprises and extends the argument of Tomkow & Vihvelin,2017 and is intended for a less philosophically specialist audience.

Relativity saying precisely what counts as the "total state of the universe at a time" is complicated and controversial [1].)

And, despite popular (because popularized) misconceptions, Quantum Mechanics is also deterministic. That is: given the total state of any system at any moment, described in the terms of Quantum Mechanics, the laws of Quantum Mechanics allow us to deduce its *quantum state* at every moment past and future.

The chanciness associated with QM arises from the fact that the quantum states of the world map only probabilistically onto classical states. Properly speaking this is not a failure of determinism. It's not that QM only gives probabilities for that cat's being alive or dead. It's that the world QM describes doesn't contain cats; it contains catty wave functions and those it predicts with complete precision. The problem ("the Measurement Problem") is that when we want to know whether the predicted quantum state corresponds to a live or dead cat, we have to start flipping coins.

This apparent mismatch of the quantum and classical worlds is puzzling, especially for those of us who think that all said and done, there is only one world. Still, we can take comfort in the fact that neither Quantum theory nor General Relativity can claim to be the whole story. Both theories can yield nonsense (or "singularities" as the physicists call them) in some limit cases and we don't yet have an account of quantum gravity. So, we expect a successor theory. Whether or not that theory will acknowledge the existence of cats, as opposed to feline wave functions or perhaps a multiversal plenitude of cats, is an open question, but we can expect that it will be, like its predecessors, deterministic.

But there is a puzzle. Determinism *seems* to contradict our commonsense understanding of the way the world works.

Determinism is a claim not just about the way the world is, but the way it *would be* if things were different. Leibniz noticed this early on:

> … ultimately, because of the connection of things, the whole universe with all its parts would be very different and would have been a different one from the beginning if the least thing in it went differently from how it does go. On account of this, it is not the case that events are necessary, but it is the case that they are certain,

[1] Cf. Earman, Hoefer, and Maudlin.

> given the choices that God made of this possible universe, whose concept contains this sequence of things. [1]

This is the *counterfactual* version of determinism[2]

> Counterfactual Determinism
>
> If the world were different from the way it is in any respect at any time, it would be different in some respect at every time.

This says that the future would be different if the present were different. But it also tells us that if the present were different, the past would be different as well. As Susskind says ,"It goes in both directions."

The puzzle is that this seems to contradict our commonsense intuition that what will happen in the future depends upon what is happening now in a way that what has happened in the past *doesn't* depend upon what is happening now.

Thus, it seems clearly true that if the Japanese had not bombed Pearl Harbor on Dec. 7, 1941, then the United States would not have declared war on Japan on Dec. 8. But if the Japanese hadn't attacked on Dec. 7, what would have happened on Dec. 6?

We think that what we will do next depends upon what we decide now, but we don't think what we did yesterday depends upon what we decide now.

This phenomenon is sometimes described[3] as "the temporal asymmetry of counterfactuals", but that label can be misleading. It is not that the counterfactual relation itself is asymmetrical, although of course, it is. From

> If Jones hadn't struck the match, it wouldn't have lit.

We cannot conclude

> If the match hadn't lit, Jones wouldn't have struck it.

In symbols:

$$A \mathrel{\Box\!\!\rightarrow} C$$

does not entail:

$$C \mathrel{\Box\!\!\rightarrow} A$$

---

[1] Leibinz. *Letters to Arnauld* , in Strickland 2007, p.44
[2] Cf. Tomkow & Vihvelin, 2016
[3] Cf. Lewis,1979

If the match hadn't lit it might have been because it was wet, or Jones might not have struck it hard enough, or the match might have been badly manufactured, or there might have been a gust of wind… .

The "asymmetry" is rather a disparity in *numbers*: there are lots of true forward-looking counterfactuals, but clear examples of temporally backward ones are hard to find.

This dearth of true backtrackers is closely tied to another important temporal asymmetry: that of causation. Again, the puzzle is not about the asymmetry of the causal relation itself but rather in the disparity of numbers. Outside of science fiction time travel stories, we don't have any examples of causal statements in which effects precede their causes. And this seems of a piece with the shortage of temporally backwards counterfactuals. Ordinarily[1], we say that some event e_1 caused an event e_2 only if we are prepared to say that e_2 would not have happened if e_1 hadn't happened. Causation and counterfactual dependence go hand in hand. If we could understand why past events don't counterfactually depend on future ones, we would be well on our way to understanding why causation has a temporal direction.

How can we square all this with Counterfactual Determinism? If any small change in the way things are now requires that things would have been different in the past as well as the future, why are there so many clear truths about the way the future would be different if the present were different but not *vice versa*?

For an answer, we need to think a bit about the logic of counterfactuals.

Begin by noticing that not all counterfactuals are created equal. Some counterfactuals are true because their antecedents are logically sufficient for their consequents. Thus:

> If there were more than ten matches in the box, then there would be more than nine matches in the box.
>
> If Jones were married, he wouldn't be a bachelor.

Call these "logical counterfactuals".

[1] Students of the subject will know the recalcitrant exceptions — preemption, over-determination, double-prevention, trumping &c.— but will, we hope, agree that such cases are not ordinary.

Then there are counterfactuals which are true because their antecedents express *nomologically* sufficient conditions for their consequents. Thus:

> If there were no oxygen present, the match would not have burned.
>
> If the match had remained at room temperature, it wouldn't have burst into flame.

Call these "nomological counterfactuals".

Since logical implication isn't temporally asymmetric, it is not surprising that we can find plenty of true logical counterfactuals that track backwards in time. For examples.

> If today were the centenary of my birth, then I would have been born a hundred years ago.
>
> If Armstrong had been the second man on the moon, someone else would have to have got there first. [1]

Similarly, given that determinism goes in both temporal directions, we should not be surprised to find temporally backtracking nomological counterfactuals. And so we do:

> If Kennedy had run for a second term in 1964, he wouldn't have been assassinated in '63.
>
> If the patient were symptomatic now, he would have to have been infected days ago.

Whenever *P* expresses a nomologically necessary condition for *Q*, ($\sim P \boxrightarrow \sim Q$) will be nomologically true.

$$\boxed{L}\,(Q \supset P) \equiv \boxed{L}\,(\sim P \boxrightarrow \sim Q)$$

Since, if the laws make it impossible that *Q* unless *P*, then it must be that if *P* were false, *Q* would also be false

In fact, determinism entails that for *every* nomologically contingent fact about any moment in time, there must be at every moment— past and future— some fact about the world at that time upon which it nomologically, counterfactually depends. To see this, take any proposition *P* which expresses a truth about the actual world at time t_0. If *P* is nomologically contingent, then there will be some nomologically possible worlds where *P* is true at t_0 and some where it is not.

[1] See Tomkow 2013 and Tomkow & Vihvelin 2017 for discussion of the significance of the "would have to have been" locution in backtrackers like these.

Take all the worlds where *P* is true at t_0 and describe the total state of each such world at some other time, t_1. Now disjoin all those descriptions to get a proposition, F_1. The disjunctive F_1 will be true at every and only those nomologically possible worlds where *P* is true at t_0. So, F_1 names a fact about the world at t_1 which is nomologically necessary and sufficient for *P* at t_0.

$$\boxed{L}\,(P \equiv F_1)$$

Q.E.D, for every fact *P* about a deterministic world at any time there must be at every other time, past or future some fact *F* about that time which is nomologically necessary and sufficient for *P*. This, in turn, entails that *P* nomologically, counterfactually depends on *F*.

$$\boxed{L}\,(\sim F \mathrel{\Box\!\!\rightarrow} \sim P)$$

So there is no shortage of nomologically true, temporally backtracking counterfactuals. In a deterministic world, there is one for every fact about every time.

However, there are counterfactuals of another sort. Consider our old friend:

If the match had been struck, it would have lit.

Said of a match that was not struck and did not light. Obviously striking a match doesn't *logically* imply that it lights. And there is no law of nature that says striking guarantees lighting. No matter how hard it was struck, the match wouldn't have lit if it was wet, or there was a strong wind, or if it was badly made…. Of course, determinism means that at any physically possible world where the match lights there will have to be a prior condition of the world which is nomologically sufficient for its lighting. We don't think the striking of the match is such a condition, but we do think that it could be *part* of such a sufficient condition. And we think, moreover, that such a sufficient condition *would* have obtained if this match had been struck here and now; because, as a matter of fact, this match *is* dry, it *is* well made, there is no wind and so on. The counterfactual is not true just because of the laws, but also because of other facts on the ground. We'll call this a "contingent counterfactual" and say that *P* contingently counterfactually depends on *Q* when *P* and *Q* are true and

$$\sim P \mathrel{\Box\!\!\rightarrow} \sim Q$$

But not

$$\boxed{\boldsymbol{L}}\,(\sim P \supset \sim Q)$$

It is among the contingent counterfactuals that we find a shortage[1], of true backtrackers.

We have a contingent counterfactual dependence whenever there is a counterfactual dependence between nomologically independent facts. It is *contingent* counterfactual dependence that pairs with causal dependence. Indeed, to a first approximation, contingent counterfactual dependence just *is* causal dependence[2].

That causation is specific to contingent dependence explains our commonsense intuitions about the difference between causes and *necessary conditions*. We know that the presence of oxygen is nomologically necessary for the lighting of matches. But no one— at least no one not in the grip of a philosophical theory[3] — would say that the presence of oxygen causes the match to light. Likewise, we do not want to say that being born causes anyone's senile dementia even though we know that you cannot have the latter without the former. We don't count these counterfactual dependencies as causal precisely because they are nomological.

This also explains the often remarked upon fact that 'cause' does not figure as a term of any scientific law. Bertrand Russell took this as evidence that causation was not a real feature of the world.[4] Rather, it is what we would expect if causation is exclusively a matter of contingent dependence. The laws expose nomological connections between goings on. Causation involves dependencies that are real but depend on local contingencies as well as the laws.

Still, we are not here to argue for a full-blown counterfactual analysis of causation. Suffice it, for present purposes, to note that if we are to understand why causation has a temporal direction we are going to have to understand the temporal directedness of contingent counterfactuals.

As our next step on that path, notice that determinism is a thesis about the *logical* properties of the laws of nature when applied to descriptions of the total state of the world at different times.

---

[1] There are *some* examples of backtracking, contingent counterfactuals which are arguably true but these seem to involve special contexts. A topical example might be, "If Hillary were president she would have to have won the electoral college". But these will this seem correct only in conversational contexts where we are treating constitutional law *as if* it were as immutable as natural laws. Cf. Bennett 2003 for a discussion of such cases.

[2] On the understanding that causation is the ancestral of causal dependence. Cf. Lewis, 1973. Vihvelin and I think that causation *can* be wholly analyzed in terms of contingent counterfactual dependence, but that is an argument we will make elsewhere.

[3] Some philosophers would insist that the presence of oxygen *must* be counted as cause of a fire just because the presence of oxygen is a necessary part of a nomologically sufficient condition for the fire. But, if that were so we would have a world full of backward causation since, as we have just observed, determinism entails that every present fact has future NS conditions.

[4] Russell,1992

Susskind calls it "reversibility" and it is not hard to see why. If we think of any physically possible world as a sequence of total physical states, <S_i, Sj, … Sn> , determinism invites us to think of the laws of nature as a function, $\mathcal{L}$, which maps total states onto total states.

$$\mathcal{L}(Sn, t) = Sn+t$$

Where t is some interval of time. This means the laws are temporally reversible insofar as t can take a negative or a positive value; symmetrically mapping future to past as well as past to future:

$$\textbf{(R)}\quad \mathcal{L}(Sn,t) = \mathcal{L}(Sn+t,-t)$$

However, calling this sort of symmetry, "reversibility" risks confusing information conservation with a quite different phenomenon: the one that Huw Price calls "T-Symmetry" and Sean Carroll calls "time reversal invariance". Here is Price:

> Roughly, this symmetry amounts to the principle that if a given physical process is permitted by physical laws, so too is the reverse process—what we would see if a film of the original process were shown in reverse. [1]
>
> … since the nineteenth century, the dominant view in physics has been that… the laws seem essentially symmetric, in the sense that any interaction which they allow to occur with one temporal orientation is also allowed to occur with the opposite orientation (the laws showing no preference between the two).[2]

This is not wrong, but it is important to distinguish this kind of "reversibility" from the kind entailed by determinism. Determinism is a claim about the connections between *total* states of the world at different moments, the kind of reversibility Price is speaking about is not.

To see this, take the stock example of T-Symmetry: a film of billiard balls bouncing around a table. The film will show the balls obeying Newton's laws whether we run the film forward or backwards and so we might rightly say that the laws allow the succession of physical states depicted on the film to occur in either direction. But notice that each frame of the film only depicts the *position* of the balls at a particular time, and that is not their *total* physical state at that time—at least not if we suppose that "total states" satisfy (R)[3]. To tell us what has happened

---

[1] Price, Huw. 1997,p.18.
[2] Ibid. p. 116
[3] There is room for terminological quarrels about what is meant by "the state of the world at a time". David Albert notes that one could maintain that Newtonian Physics *was* reversible over *total* states if one insisted that-- because velocity is really a kind of cross-temporal property-- it shouldn't be included in what is properly called "the state of the world *at a time*". Which is all very well, except that now what will be called "the state of the world at a time"

before or after any frame on the film, Newton needs to know the position and velocity of every ball. Now we can determine the balls' velocities by watching the film— inferring the velocity of each ball from its change in position from frame to frame— but reversing the direction of the film will display the velocity of the balls as reversed. So, reversing the film does *not* show us the same *total* states in reverse order, it displays every ball as having a *different* momentum at every moment.

As Carroll says:

> The lesson of all this is that the statement "this theory is invariant under time reversal" does not, in common parlance, mean "you can reverse the direction of time and the theory is just as good." It means something like "you can transform the state at every time in some simple way, and then reverse the direction of time, and the theory is just as good."[1]

So to say that physical theories obey T-symmetry is *not* to say that if the laws allow a sequence of total states, <S_i,S_j..S_n> they must also permit those states to occur in reverse temporal order, <S_n…. S_j,S_i>. No physical theory does that. Rather, it is to say that each theory provides some (simple) transformation function $\boldsymbol{f}(S)$ from total states onto total states such that if <Si,Sj…Sn> is physically possible then <$\boldsymbol{f}(S_n)$….$\boldsymbol{f}(S_j)$, $\boldsymbol{f}(S_i)$> is also physically possible. In Newtonian mechanics, the relevant function reverses velocity; in electro-dynamics, it reverses the direction of the magnetic field; in quantum mechanics, it gives the complex conjugate of the wave function.

What is important to note for our purposes is that the thesis of determinism and the idea that physical theories are T-symmetric are different and logically independent. [2] Laws might be T-symmetric without being deterministic and *vice versa*.[3] The kind of reversibility that determinism requires and (**R**) asserts is *logical*. It is about making backwards *inferences*, not making things run)backwards.

A logically reversible (or "invertible") function is one whose arguments can be inferred from its values; if we know what it outputs, we can infer its inputs. Mathematical negation, -x and

---

won't include enough information for physics to tell us about past or future states so these cannot be the "total states" that the determinist is talking about.

[1] Carroll, 2016. p. 134.

[2] Alas not all physicists are as clear as Carroll about the difference between time invariance and deterministic reversibility. Susskind, for example, often conflates the two. Cf. Susskind, 2015

[3] See Albert, 2000, ch.1 for examples

logical negation, ~p are the paradigm reversible functions. If you know the value of $-x$ is 3, you know x is -3. If $\sim p$ is false, then p must be true. Addition is not reversible; you can't infer the value of x or y from the fact that x +y = 5. Some functions are reversible from some outputs but not others: if $(p \,\&\, q)$ is true both p and q must be true, but if it is false it might be that p is false, or q or both. And some functions are reversible when applied to some domains but not others: x^2 is reversible when x is restricted to positive numbers, but not when it operates over real numbers.

Now notice that Determinism requires the laws of nature to be reversible only when they operate over *total* states—complete descriptions of all the facts about the world at a time. It is consistent with determinism that knowing only *some* of the facts about time t_1 might allow you to infer some of the facts about t_2 but that knowing those t_2 facts would not allow you to say how things were at t_1. And it is consistent with determinism that the laws might imply a good deal about the future given a few facts about the present but entail little about the past given only those present facts.

While this point should be obvious, it is sometimes overlooked. Here, for example, is Stephen Hawking dismissing Reichenbach's view that the direction of time is the direction of causation:

> [Reichenbach] laid great stress on causation, in distinguishing the forward direction of time from the backward direction. But in physics, we believe that there are laws that determine the evolution of the universe uniquely. So, if state A evolved into state B, one could say that A caused B. But one could equally well look at it in the other direction of time, and say that B caused A. So causality does not define a direction of time. [1]

This is wrong on two counts. First, the fact that the total state of the universe at A uniquely determines a future state B should *not* lead us to say that A causes B otherwise we would indeed be obliged to say that B causes A. That we are not tempted to say that the future states of the universe cause its past states shows that we do not treat nomological necessity as the same thing as causation and hence that there is no contradiction in saying one has a direction and the other does not. The second point is that determinism does not tell us that physical state A must always lead to state B and vice versa if states A and B are anything *less* than total descriptions of the state of the universe And, as a matter of fact, the states of the world over which we make causal

[1] Hawking 1994

and counterfactual claims— the states that scientists observe, measure and predict, the states that serve as arguments for actual laws— are invariably *less* than total.

Take the paradigm deterministic system: a Newtonian universe of elastically colliding billiard balls. Of course, given the position and momentum of every ball at any given time, Newton's laws entail the total state of the system at every other time. But Newton's laws are themselves not logically reversible when they operate over less than total descriptions. Given that a ball has an initial velocity, v_i, and is subject to an average acceleration a for an interval t we can calculate its final velocity as:

$$v_f = at + v_i$$

But given only the information that the ball's final velocity is v_f we cannot deduce its initial velocity, the magnitude of a or the length of t. And notice that this inferential asymmetry is wholly consistent with T-symmetry: if we reverse the velocities we can "reverse the film" of the balls motions, but we still will not be able to deduce the ball's new v_i from its new v*f*.

Not every natural law is irreversible in this way. Conservation laws tell us precisely that if some quantity has a certain value at one time, then it will have it later and sooner. But the *dynamic* laws of every physical theory are irreversible[1]: thus the time-dependent Schrödinger equation allows us to determine that $\psi_i(x, 0)$ will evolve to $\psi_f(x,t)$ given an energy of $\mathcal{H}$ for an interval t. It does not allow us to determine $\mathcal{H}$ and $\psi_i(x,0)$ given only $\psi_f(x,t)$

A logically irreversible operation represents a loss of information about the system's initial state. The final state of the ball or the wave function at t_f does not tell us everything about its initial state. This failure of *local* determinism for partial states is perfectly consistent with *global* determinism over total states. We can't know everything about the history of the ball just from its current position and momentum, but if we take account of the current position and momentum of other balls we can build a picture of which balls collided with which, and when, and piece together the whole story. The lost information about the history of one ball is preserved in the current facts about other balls. Still, these local, logical asymmetries *do* have consequences for the counterfactual relations between the states they connect.

[1] Indeed, irreversibility is the clearest criteria for drawing the often ill-drawn distinction between dynamic and kinematic laws.

It is the irreversibility of laws that makes *contingent* counterfactual dependence possible. To see this, suppose that p_1 is a fact about t_1 and q_2 a fact about t_2 and

$$\sim p_1 \boxdot\!\!\rightarrow \sim q_2$$

For this to be so in a deterministic world, it must be that p_1 is entailed by some fact r_1 about t_1 such that[1]:

$$\boxed{L}\,(r_1 \supset q_2)$$

But now suppose that the relation between r_1 and q_2 is reversible so that:

$$\boxed{L}\,(r_1 \equiv q_2)$$

Given that r_1 entails p_1, we have:

$$\boxed{L}\,(\sim p_1 \supset \sim q_2)$$

Which would make $\sim p \boxdot\!\!\rightarrow \sim q$ nomologically necessary.

We find contingent dependence whenever we find states connected by irreversible laws. If the initial v_i of our ball at t_0 had been 1m/s less then its v_f would have been 1*m/s* less though the laws don't make an initial velocity of v_i *necessary* for having a final velocity of v_f. The counterfactual is true not just because of the laws but also because of other facts about the world that obtain at t_0.

The connection between logical irreversibility and counterfactual asymmetry is less obvious. As we have observed, a local failure of reversibility over partial states is consistent with global reversibility over total states. When we conjoin the fact that the velocity of the ball at t_0 was v_i with everything else that is true about the world at t_0, we will be able to deduce that the ball will have a final velocity of v_f. And, as deterministic reversibility requires, when we combine the fact that the ball has v_f at t_1 with everything else that is true about the world at t_1, we will be able to deduce that the ball had v_i at t_0. But this global deductive symmetry does not give us counterfactual symmetry. If v_i had been 1*m/s* less than it was at t_0, v_f would also have been 1*m/s*

[1] Given determinism, the total state of the world at $t_1$must be one such *r1* fact, but for the sake of this argument we need not assume *r1* is total.

less at t_f, but, if v_f had been 1*m/s* less at t_1, v_i would not have to be 1 m/s less at t_0; it might have been, instead, that *a* was less.

We might ask then: "If we can deduce v_i at t_0 from the conjunction of v_f with all the other facts about the world at t_1, why can't we just add 1 *m/s* to v_f, conjoin that with those other facts and deduce what v_i would have been"? The answer is that we can't do things this way because this conjunction would describe a physically *impossible* world state. If v_f wasn't what it was at t_1 then the laws will require that some other facts about the world at t_1 would also have to be different.

To see why this must always be so, remember that one-way functions lose information about their inputs. If information is to be conserved globally, then the lost information must be preserved elsewhere. The t_1 position and momentum of a single billiard ball tells us little about its past. To fill out the story we must look at the current position and momentum of all the other balls and reconstruct the history of their movements and collisions. All the information about the balls position and momentum at t_0 is preserved at t_1, but it has been, in effect, dispersed among all the other balls that might have collided with it in the interval. If the ball's current velocity were different, then it would have to have had a different history, and that would have to be reflected in the current state of other balls. To use the philosopher's jargon: the ball's having a different velocity than it actually does at any given moment is not *co-tenable* with the actual facts about all the other balls. Two propositions are not co-tenable if one would be false if the other were true. [1]

For a ball's velocity to be different than it is *now*, then other things would have to be different *now*. And precisely because its current velocity is an irreversible function of multiple factors, if that velocity were different, there is no one way things would have to have been different in the past and, hence, no one way things would have to be different in the present.

To say what would have happened in the past or the future if the balls current velocity was different than it actually is, we must choose among the many different ways the *present* might have to be different if that were the case. After excluding the physically impossible alternatives,

[1] Goodman, 1947. p 120

we decide what would have happened by examining the alternative co-tenable situations which are *minimally* different from the way things actually are.

More abstractly, the recipe is this:

(C) Where *p* is a truth about the world at time t_x and *q* is a truth about the world at time t_y then ~p $\Box\!\!\rightarrow$~q if and only if there are total states where ~*p* & ~*q* that are more similar overall to the way things actually are at t_x than in any total state where ~*p* & *q*. [1]

A digression: The fact that the logic of counterfactuals rests their truth conditions on this sort of similarity metric has led some authors to view them as suspect. Who is to decide what differences are "minimal" or what makes one counterfactual situation more like reality than some other? Aren't such judgments subjective and context dependent? The great philosopher W.V.O. Quine captured[2] the slipperiness of some counterfactual talk with the examples.

> If Julius Caesar had been in command in the Korean War, he would have used nuclear weapons.
>
> If Julius Caesar had been in command in the Korean War, he would have used catapults.

There is obviously no right answer to which of these is correct because counterfactual situations with a later Caesar or an earlier Korean War must both be so wildly different from the actual world that there is no single sensible standard to judge which is more similar. Depending on our interests and the context of a given conversation we might find ourselves endorsing one or the other.

But it does not follow from these quotidian cases that all is flux. Ultimately, the truth of counterfactuals always depends on the facts and the laws. We can say what happens next in any counterfactual situation only if we know what laws of nature operate there and how the facts are different there. The ambiguities in Quine's example come about because it is not clear exactly what the relevant counterfactual situation is supposed to be like in ways that are nomologically relevant to what comes next.

The more precisely the antecedents of a counterfactual describe the world in the language of physical theory, the less ambiguous its truth conditions become. This is why counterfactual talk

---

[1] Notwithstanding their other differences. Principle (**C**) is held in common by every account of the logic of counterfactuals including Lewis, Bennett and Stalnaker. Cf. Tomkow,2013.

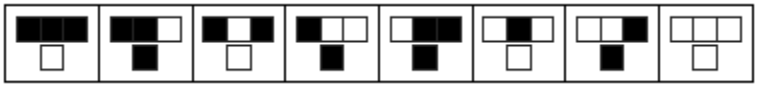
[2] Quine, 1960. p.222

is central to science. “If the energy density of the vacuum were greater than it is, the universe would be expanding more rapidly.” may or may not be true but its truth is not a matter of arbitrary convention or context.

To return to our governing concern: Our formula (**C**) for evaluating counterfactuals does not assume any sort of temporal direction. It applies whether t_x precedes follows or is simultaneous with t_y. And yet if we apply this standard to a world which is globally deterministic but locally indeterministic a direction emerges.

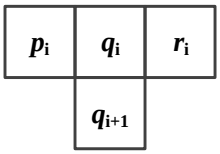
To see this, we can try out these ideas with a computer simulation of a deterministic universe with logically irreversible laws. It is a simulation we can run on paper.

A one-dimensional, binary cellular automaton (CA) is a string of adjacent elements (“cells”) each of which can take only one of two states. The cells can switch states over time. Whether any given cell will switch depends on its state and the state of its immediate neighbours. The rules describing these dependencies define the CA.[1] Here is Stephen Wolfram’s Rule 90 in black and white.



Rule 90

In a rule 90 CA, a cell’s state is entirely determined by the previous state of its two neighbouring cells. If they are the same colour, it will be black; if they are different colours, it will be white. If we read black and white as true and false then, if p_i, q_i, r_i are adjacent cells,



we can express rule 90 as the “law”:

$$(\text{L}90) \quad q_{i+1} \equiv (p_i \oplus r_i)$$

L90 is logically irreversible. We cannot infer the values of p_i or r_i given value of q_{i+1}.

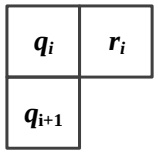
[1] Cf. Wolfram 2002 for a comprehensive discussion. Francesco and Tagliabue, 2017, offer an excellent overview.

Though simple, Rule 90 can produce results of considerable complexity. Figure 1 shows two examples of the rule applied to a row of 50 cells.

Figure 1[1]

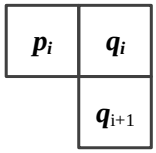
The figure on the left shows the results of beginning with a single true cell. The one on the right, with a random initial condition. The cells at the boundaries are treated as if the cell beyond the limit was false. So, if q is the cell at the far left (L) reduces to:

(L90') $q_{i+1} \equiv r_i$



and for the rightmost cell:

(L90'') $q_{i+1} \equiv p_i$



These boundary laws are logically reversible.

With this "null boundary" condition in place, any finite Rule 90 CA with an even number of cells is globally reversible[2]. That is: given the state of every cell at any time, we can deduce the total state of every cell at every time, notwithstanding the fact that Rule 90 itself is irreversible.

As we should now expect, this combination of local indeterminism with global determinism has consequences for what counterfactuals are true in these model universes. There are nomological counterfactuals going forward and backwards. [3]

---

[1] Figure 1 from Rey,2015. Generated by Mathematica.
[2] Cf. Rey 2015.
[3] CA-90 could also be said to be time reversal invariant though it is such a simple function that there are several candidates for its inverse. I suggest the simplest inversion would be negation, so that qi+1 ≡ ~(pi ⊕ ri). This

$$\boxed{L}\,(\sim(p_i \oplus r_i) \Box\!\!\rightarrow \sim q_{i+1})$$

$$\boxed{L}\,(\sim q_{i+1} \Box\!\!\rightarrow \sim(p_i \oplus r_i))$$

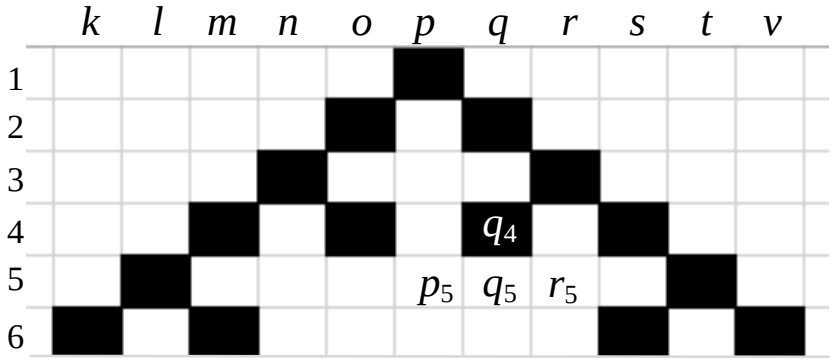
And for our special boundary cells (L90') gives us:

$$\boxed{L}\ ((\sim q_{i+1} \Box\!\!\rightarrow \sim r_i) \ \&\ (\sim r_i \Box\!\!\rightarrow q_{i+1}))$$

And (L90")

$$\boxed{L}\ ((\sim q_{i+1} \Box\!\!\rightarrow \sim p_i) \ \&\ (\sim r_i \Box\!\!\rightarrow p_{i+1}))$$

To see how things play out for the non-boundary cells, consider this close up of one of our Rule 90 worlds.



CA-90

How would it look if q4 were false? We cannot just invert q4 and apply L90, leaving all the other cells as they are. There is no rule-90, legal history that could lead to that pattern at t4. Because it is deterministic, any change in any cell in CA-90 at any time will require changes in the past, and because L90 is irreversible, those changes must affect other cells in the present as well as the future. There is no lawful way to change *any* single cell in any row of this CA without supposing that at least one other cell in that row would be different. [1]

would correspond to Wolfram's Rule 165. The fact that rule 165 CA's do not look at all like rule 90 CAs "running backwards" only highlights the difference between T-symmetry and deterministic reversibility.

[1] The exception is Row 2 where alternatives *can* be created with only a 1-cell difference if we allow row 1—our initial condition—to contain "illegal" configurations. That is, patterns that could not have legally evolve from any other pattern. Whether or not our simulation would more closely mirror reality if it allowed such "singularities" in

That this is so—that a single cell difference is not co-tenable with everything else remaining the same— is a necessary upshot of combining global determinism with irreversible laws. Transitions governed by L90 lose information about the prior state of the cell and its neighbours, for that information to be preserved going forward it must be reflected in the state of other cells. L90 entails that any difference in any cell will require at least two cells to be different in the next row.[1]

That number, two, is significant because it represents the *minimum* legal change that can be made to any row at any time. That means that if we ask what would happen if any particular cell were different at a particular time, we must consider counterfactual situations that are different in at least one other cell at that time. There will be many such situations corresponding to different histories with different initial conditions. But, it so happens[2] that of all the possible histories that would have brought it about that ~q4, those that are *minimally* (2 cells) different from CA-90 at t4— though they all differ from one another — have in common that r_5 and p_5 are true in the next generation. So, applying (**C**) we conclude:

$$\sim q_4 \Box\!\!\rightarrow r_5$$

$$\sim q_4 \Box\!\!\rightarrow p_5$$

But among this equivalence class of most similar worlds, q5 is true and some but not at others. So:

$$\sim q_4 \Box\!\!\rightarrow q_5$$

is *not* true at CA-90.

These counterfactuals are all contingent. They have the values that they do only because of the values that other cells happen to have in CA-90 at t_4.

But now what about the other temporal direction? If q4 were false how would things have been different earlier on? The answer is that there are no true, *contingent* backtracking

---

its initial conditions is a nice question but does not affect anything that follows. There is no counterfactual backtracking from t2 to $t_1$even with illegal initial conditions.

[1] Again, with the possible exception of Row 2.

[2] I don't say it is obvious that this is so. It is part of the peculiar "science" of complex systems that there is often no non-brute force way to prove such claims. I am reporting the results of countless automated CA runs. Readers are welcome to test them against their own simulations.

counterfactuals in CA-90. Any legal history which results in q4 being false must also result in some other cell being false. But of all the minimally different legal histories that result in q4 being false (and there are many such) there is no single cell which has a different value than it actually does in every such history. Of course, if q4 were false, then its neighbours would have had to have different colours at t3.

$$\sim q4 \mathrel{\Box\!\!\rightarrow} \sim(q3 \oplus r3)$$

But that counterfactual is nomologically true, not contingent.

At the boundary, we will be able to say that, e.g. that if the left-most cell were reversed at t4, then the left-most-but-one cell would have to have been different at t3. But that counterfactual too is nomological given (L90').

In CA-90, as with our billiard balls, contingent counter-factual dependence always goes forward. If we were to map out the counterfactual dependencies of a single cell the results would look like figure 2.

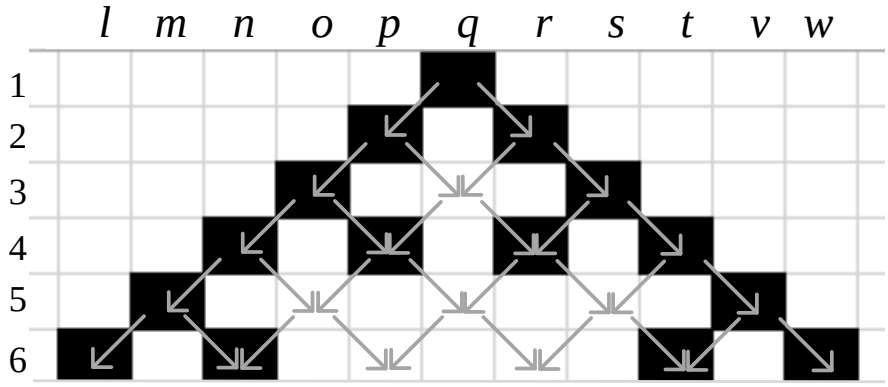


Fig 2

A complete picture would show parallel cascades of dependence moving forward from every initial cell. Notice too that the picture doesn't just illustrate counterfactual dependencies between cells at different times but implies counterfactual dependencies amongst cells at the same time. And these dependencies, like the counterfactual influence of a single cell, spread over time. At t_2 the state of cell q can make no difference to the state of cells p or r. By t4, both p4 and r4 counterfactually depend on q3.

This map of dependencies also tracks the flow of information amongst the cells as they evolve through time. The state of every cell represents 1 bit of information. Each cell passes forward

information about its state to its neighbour's successive states. Cell p2 carries the information that q1 and o1 had different colours at t_1. Cell r1 carries the information that q1 and s1 were different colours at t_1. Each t_2 cell has got one bit of information about the prior states of its two neighbours. Together p2 and r2 have two bits of information about the states of three cells at t_1: o1,q1 and s1. That leaves one bit of information about t_1 missing: the information that would decide if q was black and the other two white or vice versa. And so it will go up and down the line till we reach the borders: every n cells at t_2 carry n-1 bits of information about t_1. The missing bit of information about t_1 is carried forward, redundantly by the boundary conditions. The state of each boundary cell losslessly records one full bit of information about a prior cell and that, given the information in all the other cells at t_2 determines the total state at t_1. The boundary conditions serve as conservation laws. By preserving the state of a single cell, they guarantee the over all conservation of information over time.

Information about q's state at t_1 is carried by the system's total state at t_2, but it is mostly concentrated in p2 and r2. As time goes on it will be more diffused. At t3, the same information has spread to o3, q3 and s3. This spread is reflected in the spread of counterfactual interdependence among all the cells at successive times. In terms of Shannon Information theory, it constitutes an increase in the *total correlation* of the cells— the averaged amount of mutual information that any given cell has about the other cells of its generation— given the past state of any cell or set of cells.

And finally, though we will not press this point. The direction of these arrows of counterfactual dependence also mirrors our commonsense intuitions about the causal connections amongst these cells. Cause has a direction because counterfactual dependence has a direction and that direction stems from the logical structure of the laws.

It is important to understand that the direction of these dependencies has nothing to do with any hidden assumptions we are making about the temporal direction of the cells' evolution. Our CA-90 system is two-way deterministic. That means that we can construct an algorithm and write a program which could— given any later, *total* state of CA-90— deduce and display the row of cells evolving backwards and ending up where our CA began. Even so, so long as what our algorithm outputs conform to Rule-90, all of the counterfactual dependencies we have noted will

be preserved. For example, it will still be true that if q4 were false, then r5 would be false (or as we would then say, would have to *have been* false) even if row 5 was printed before row 4.

We can reverse the temporal order of the total states of our CA-90 because it is only a simulated world and rule-90 is not really a law of nature. In the real world, the real laws do not allow total states to occur in reverse order nor can we invert the counterfactual and causal dependencies the laws impose.

Here again, though, we must be careful to distinguish reversing the temporal order and T-Symmetry.

We never observe the total state of the universe at any time. We observe bits of it. A cup falls off a table and shatters on the floor. A rock drops in a pond and concentric rings of waves radiate outward. It seems we always find these sorts of things happening in a certain order. But the time invariance of physical theory means that it need not be so. If the molecules that constitute those shards on the floor had just the right momenta, they could leap off the floor and reassemble the cup. Unlikely though it might be, if the random kinetic forces latent in the bank of the pond were suddenly to align, they could send concentric waves rushing to the centre of the pool and eject that rock. Even so, in such cases, the direction of counterfactual and causal dependence is always from past to future. If that shard had not been precisely where it was, the assembled cup would have lacked a handle. If the river bank had been a different shape, the centring wave would not have been circular.

Hue Price disagrees. Price argues that if we set aside the distractions of kitchens and ponds and focus on the underlying physics, we find that all asymmetry disappears.

> … consider a simple and clearly symmetric physical example, such as the collision of two (frictionless) Newtonian billiard balls. If we know the combined momentum of the balls before they collide, then conservation of momentum ensures that we also know it after the collision. After the collision, then, the momentum of the balls is correlated, in the sense that by measuring the momentum of one ball, we can determine that of the other.
>
> … however, there is nothing asymmetric about this. After all, if we know the combined momentum after the collision (which is the same thing as knowing the momentum before the collision, of course, thanks to conservation of momentum), then we know that the momentum of the balls before the collision is correlated in just

> the same way. (The fact that there is nothing intrinsically asymmetric in cases like this was the point of Figure 6.2.)[1]

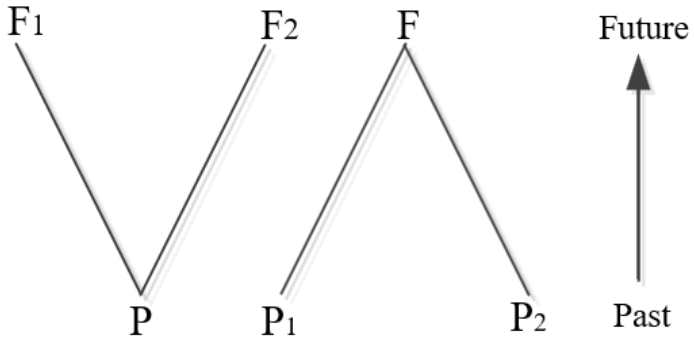


Figure 6.2

Price explains.

> Suppose we consider micro events P, F1, and F2, as in Figure 6.2. It will typically be the case that if P is held fixed, then F1 and F2 may be correlated with one another; if F1 had been different, F2 would have been different. Given conservation of momentum, for example, holding *P* fixed may ensure that the momentum of F1 is related to that of F2. But … precisely the same obtains in reverse. If we hold *F* fixed, then the same will apply to P1 and P2. Temporal direction is irrelevant.[2]

On the strength of this argument, Price concludes that at least at the level of micro-physics — where interactions can be isolated from exogenous factors like friction— there is no counterfactual or causal asymmetry. He doesn't think this shows that micro-physical causation can go backward, but rather that the direction of counterfactual dependence and causation is not an objective feature of the world.[3]

Seeing what's wrong with this will be a useful way to reprise points already made. For a start, let us put a little flesh on Price's extremely schematic diagram. Here is the sort of situation he is describing:

[1] Price 1997,p. 181.
[2] Ibid. p. 150.
[3] Ibid. 1997 ch. 6.

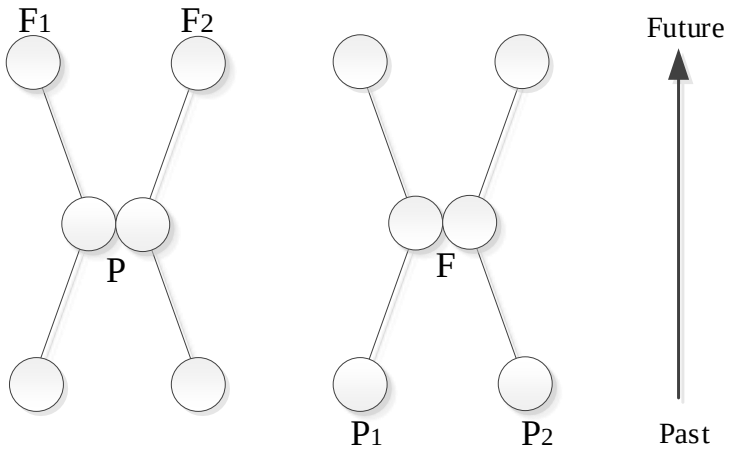


Figure 6.2*

The first point to make is that to invoke conservation of momentum here requires treating these goings on as if they were closed systems. That, in turn, means treating them as locally deterministic.

The second point is that there is symmetry here precisely because Conservation of Momentum *is* a conservation law and conservation laws are logically reversible.

The third point, which follows directly from the second, is that conservation laws do not sustain *contingent* counterfactual or causal claims. If we assume that the total momentum of the system must preserve its actual value, then it does indeed follow that any change in the momentum in F1 must be reflected in a change in F2. But that inference, and the counterfactuals that follow from it are nomologically necessary. Moreover, these dependencies are unconnected with the causal facts about the situation. If we assume that the total momentum cannot vary, the momentum of F1 must correlate with the momentum of F2 *whatever* their causal history and whether or not the balls ever collided.

The fourth point is that while it is perfectly intelligible to ask what would happen if things were otherwise on the assumption that total momentum is unchanged, we do not make that assumption when we reason counterfactually about the behaviour of particular particles or billiard balls. It is plainly false that if the momentum of a ball were different at P1 or F1 that the momentum of the system would remain the same. What is true is that if the momentum of the ball at F1 was different the total momentum of the system would be

different than it actually was. Conservation of momentum says that total momentum doesn't change. It does not say that it could not be different from what it actually is.

As soon as we try to explain local changes within this closed system—e.g. the changes in the momentum of a particular particle or ball—contingent counterfactual asymmetries emerge because logically irreversible dynamic laws come in to play. Thus, if the velocity of one of the balls were different at the time of *P*, then the velocity of both would have to be different at F1 and F2. Given their history, a change in momentum at F1 is not cotenable with F2's remaining the same. On the other hand, if the velocity of one of the balls in F were different, it's not the case that both P1 and P2 would have to be different. Or again, if the collision at P had not occurred then the momentum of both F1 and F2 would be different. But if F didn't happen, at least one of P1 or P2 might have remained the same. Then too, since we are engaged here in *counterfactual* reasoning, we are allowed to ask what would happen if, for example, a third ball had cancelled P1's momentum. In that case, *F* wouldn't have happened. In contrast, if the F1 ball's momentum were cancelled any time after P, P would still have occurred. These are all contingent counterfactuals, and they reflect the fact— which was, in any case, obvious as soon as we revealed 6.2 as 6.2*— that P causes F1 and F2. F doesn't cause P1 and P2.

What is true in Price's micro-worlds is true of the world at large. Begin a whole universe with as many particles as you like distributed in any way you like. Assign them properties in any physically possible combination you prefer so long as what happens thereafter obeys the laws of nature as we understand them. Initially, the state of each particle is up to you and need not affect your assignments to the others. But as time goes on the state of each must depend on its history of interaction with the others. Eventually, the state of every particle will counterfactually depend on the past states of every other so that no state could be otherwise at any time were not other states different at that time. This spreading counterfactual dependence will go temporally forward whatever else might wax and wane. If you start your world with low entropy, entropy will probably increase: temperature will move to equilibrium; macro structures will disperse. Given enough time though, that tide will likely reverse. Chance disequilibria may form, fragmented cups may leap on to shelves, waves may throw rocks out of pools. Still, through all this, the underlying processes — countless instances of diagram 6.2*— will have the same

counterfactual direction. Which is why, through all the entropic back and forth, time will continue to move forward.

The direction of contingent counterfactual dependence is the direction of time.

In information theoretic terms, we can think of the history of the universe as an information channel carrying a message from the past to the future. The content of the message does not change, and in a deterministic universe, it is conveyed losslessly from moment to moment. The content of that message is the complete story of universe stretching, possibly infinitely into the future and the past. The whole of that story about all moments is compressed and recorded in the state of the universe at each moment. The laws of nature can decompress this momentary data to reveal the whole story. Though the state of the world at any moment tells the same story as every other, each moment tells it differently. Irreversible laws divide the information about each present micro-state and disperse it amongst multiple future micro-states. Global information is preserved, but it is also transformed. At one moment in our universe, its past and future history were encoded in a grapefruit-sized ball of quark-gluon plasma. Now the same story, including the record of that early moment, is told by trillions of galaxies scattered across billions of parsecs. How the story is told depends on when it is told, and that dependence is asymmetrical because the future depends on the past in a way the past does not depend on the future. Our labour here has been to show how that can be so.

Terrance Tomkow

Los Angeles 2018

REFERENCES

Albert, D. Z. 2000. *Time and Chance*. Cambridge: Harvard University Press. Kindle Edition

Bennett, J. F. (2003). A guide to conditionals. New York: Oxford University Press.

Carroll, S. 2016. *From Eternity to Here*. New York: Dutton. Kindle Edition

Earman, J. 1986. *A Primer on Determinism*. Dordrecht: D. Reidel.

Hoefer, C. 2003, January 23. *Causal Determinism*. Retrieved December 01, 2017, from https://plato.stanford.edu/entries/determinism-causal/

Berto, Francesco and Tagliabue, Jacopo, "Cellular Automata", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2017/entries/cellular-automata/>.

Lewis, D.K. 1973 , "Causation," Journal of Philosophy, 70: 556–567.

- 1979, "Counterfactual Dependence and Time's Arrow," Noûs, 13: 455–476.

Leibniz, G. W., & Strickland, L. 2007. *The shorter Leibniz texts: a collection of new translations*. London: Continuum.

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

Price, H. 1997. *Time's Arrow & Archimedes Point*. New York: Oxford University Press.

Quine, W.V.O. 1960. *Word and Object*, Cambridge, Mass.: M.I.T. *Press*.

Rey, Ángel. (2015). A note on the reversibility of the elementary cellular automaton with rule number 90. Revista de la Union Matematica Argentina. 56. 107-125.

Russell, Bertrand, 1992, "On the Notion of Cause," orig. 1912, in J. Slater (ed.), *The Collected Papers of Bertrand Russell v6: Logical and Philosophical Papers 1909–1913*, London: Routledge Press, pp. 193–210.

Susskind, Leonard,2007 Lecture 1 | Modern Physics: Classical Mechanics (stanford) Stanford - https://www.youtube.com/watch?v=pyX8kQ-JzHI&list=PLQrxduI9Pds1fm91Dmn8x1lo-O_kpZGk8

- 2009. *The black hole war: My battle with Stephen Hawking to make the world safe for quantum mechanics*. New York, NY: Back Bay Books. Kindle Edition
- 2015,*Boltzmann and the Arrow Of Time: A Recent Perspective* http://www.cornell.edu/video/leonard-susskind-1-boltzmann-and-the-arrow-of-time

Tomkow,Terrance. 2013. *The Simple Theory of Counterfactuals*. tomkow.typepad.com/tomkowcom/2013/07/the-simple-theory-of-counterfactuals.html

Tomkow, Terrance and Vihvelin, Kadri. 2016, *Determinism*, https://vihvelin.typepad.com/vihvelincom/2016/07/determinism.html

- The Temporal Asymmetry of counterfactuals, https://vihvelin.typepad.com/vihvelincom/2017/12/the-temporal-asymmetry-of-counterfactuals.html

Wolfram, S., 2002, *A New Kind of Science*, Champaign, IL: Wolfram Media.