Abstract for “The Consequence Argument Meets the Mentaculus”

Barry Loewer Rutgers University

forthcoming in *Time’s Arrow and the Probability Structure of the World*, ed. Loewer, Weslake, Winsberg, Harvard University Press.

The “Consequence Argument” has spawned an enormous literature in response.[1] The most notable of these is David Lewis’ based on his account of counterfactuals. My excuse for adding to this literature is while Lewis’ diagnosis of the argument is on the right track the account of counterfactuals he relies on to rebut the argument is, as I will argue, defective. I will develop a response that is in some ways similar to Lewis’ but differs in that it is based on a different and better account of counterfactuals which itself is based on an approach to statistical mechanics that goes back to Boltzmann and has more recently been developed by David Albert in his book *Time and Chance*. This account, which Albert and I refer to as “the Mentaculus”, provides a framework for explaining and connecting the various so called “arrows of time” including those of thermodynamics, causation, knowledge, and influence. It is the last of these arrows that is key to my response to the consequence argument.[2] If my response is effective, then it will turn out that physics (together with some philosophy) rather than conflicting with freedom is able to rescue it, at least, from the Consequence Argument.

[1] Van Inwagen (1983) is the *locus classicus* for the argument. Van Inwagen subsequently reformulated in various ways in response to criticisms.

[2] The response to the Consequence Argument I will defend is novel but has similarities to discussions by Hoefer, (2002), Ismael (2018, Loewer (2007) Vihvelin (2017), Beebee (2013) and Dorr (2016).

Forthcoming in *Time's Arrow and the Probability Structure of the World*, ed. Loewer, Weslake, Winsberg, Harvard University Press.

The Consequence Argument Meets the Mentaculus

> "it is as impossible for the subtlest philosophy as for the commonest reasoning to argue freedom away. Philosophy must therefore assume that no true contradiction will be found between freedom and natural necessity in the same human actions, for it cannot give up the idea of nature any more than that of freedom. Hence even if we should never be able to conceive how freedom is possible, at least this apparent contradiction must be convincingly eradicated. For if the thought of freedom contradicts itself or nature ... it would have to be surrendered in competition with natural necessity." (Immanuel Kant, Fundamental Principles of the Metaphysics of Morals, 75-6)

Kant feared that freedom may be in conflict with the "natural necessities" determined by the fundamental laws of physics. This fear, which is shared by many, has been supported by well-known philosophical arguments that aim to show that if all events including our decisions and actions conform to deterministic laws then none of our decisions or actions are freely chosen. Chief among these is the

Consequence Argument. Its basic thrust is old but its more recent formulation and defense by Peter van Inwagen spawned an enormous literature in response.[3] The most notable of these is David Lewis' based on his account of counterfactuals. My excuse for adding to this literature is while Lewis' diagnosis of the argument is on the right track the account of counterfactuals he relies on to rebut the argument is, as I will argue, defective. I will develop a response that is in some ways similar to Lewis' but differs in that it is based on a different and better account of counterfactuals which itself is based on an approach to statistical mechanics that goes back to Boltzmann and has more recently been developed by David Albert in his book *Time and Chance*. This account, which Albert and I refer to as "the Mentaculus", provides a framework for explaining and connecting the various so called "arrows of time" including those of thermodynamics, causation, knowledge, and influence. It is the last of these arrows that is key to my response to the consequence argument.[4] If my response is effective, then it will turn out that physics (together with some philosophy) rather than conflicting with freedom is able to rescue it, at least, from the Consequence Argument.

A preliminary version of the Consequence Argument goes like this:

1) The past is not up to me

2) The fundamental laws are not up to me

3) Determinism: The laws and past entail the future

Therefore

4) The future is not up to me

The upshot of the argument is that if determinism is true then the future is not up to me; since my exercise of free will requires that the future at least to some extent is to up to me it follows that I lack free will. Compatibilists about free will reject this conclusion. But to be convincing they need to say what is wrong with the argument. That is the primary aim of this paper.

Determinism means that the fundamental dynamical laws and the entire fundamental state of the universe at any time t entail the state at every other time. In classical mechanics the state of the world at t consists of the positions (or relative positions) and the momenta of all the particles and the laws are the dynamical laws of classical mechanics (e.g. Hamilton's equations). Since the position of my hands at t1 supervenes on the state of the universe at that time it follows that the state of the universe at times before my birth entail the positions of my hands at t1. This seems to mean that the position of my hands at t are not up me now since it had already been "decided" by the world's physical state and its laws before my birth. The consequence argument is supposed to make this reasoning persuasive.

In discussions of free will it is often pointed out that classical mechanics has been superseded by quantum mechanics as the fundamental theory of the world and since quantum mechanical predictions

[3] Van Inwagen (1983) is the *locus classicus* for the argument. Van Inwagen subsequently reformulated in various ways in response to criticisms.

[4] The response to the Consequence Argument I will defend is novel but has similarities to discussions by Hoefer, (2002), Ismael (2018, Loewer (2007) Vihvelin (2017), Beebee (2013) and Dorr (2016).

are probabilistic the theory is not deterministic and so conflict between physics and freewill is moot. But this is not so. While quantum theory as presented in physics textbooks is a fantastically accurate instrument for making probabilistic predictions its ontology and the dynamics that ontology obeys are notoriously obscure. The usual textbook account (the so-called Copenhagen interpretation) makes reference to "measurement collapsing the wave function" in formulating its laws without a clear account either of measurement nor the ontological status of the wave function. There are both deterministic and indeterministic interpretations or reformulations of quantum mechanics that address these issues.[5] Since I don't want to get entangled in these issues here and or the existence of free will to be hostage to the interpretations of quantum mechanics I will restrict attention to the argument that claims to show that free will is incompatible with determinism and pretend that classical mechanics is true. [6]

A decision or action is up to me only if I am able to choose among more than one incompatible action that are in some sense open or possible for me. There are differing views about the pertinent sense of "possibility" at issue in this ability. If one thinks that that the relevant notion of possibility is physical possibility at t -i.e. compatibility with the state at a time t prior to the decision and the laws- then the incompatibility of free will and determinism is immediate. No further argument is needed since there is only one choice that is physically possible at t. But there are other notions of possibility that have been proposed in accounts of the ability involved in free choice. For example, Christian List suggests that the relevant notion of possibility is compatibility not with the microscopic state and laws but with what he calls "the agential state" and whatever laws govern an agent's psychology.[7] Since it is plausible that an agent's psychological state and even her psychological state together with the macroscopic physical state of her body and its environment is not linked by deterministic law with her subsequent decisions and actions there may be more than one action that is agent-possible at t.[8] Van Inwagen's argument is aimed at showing that even if one grants to the compatibilist that in some sense an agent is able to choose among alternative decisions free will and determinism are incompatible.

My initial formulation of the Consequence Argument employed the expression "up to me." It is not clear what it takes for a state of affairs to be "up to me." I understand it to mean at least that there are more than one alternative decisions that I am able to make (i.e. are possible for me) such that my decisions *influence* which alternative states of affairs actually occur.[9] For my decisions to influence what

[5] Bohmian mechanics and the "many worlds" understanding of quantum theory both have dynamics that are deterministic. There are theories that are modifications of quantum mechanics -so called "spontaneous collapse theories" like GRW that do have indeterministic dynamics. While these theories are not subject to the Consequence argument they pose other challenges for free will.

[6] There are arguments analogous to the Consequence Argument that are claimed to show that free will is also incompatible with probabilistic theories like GRW that cover all physical events (see Loewer 1997) and van Inwagen's "rollback argument" is designed to show that indeterminism is also incompatible with free will.

[7] C. List (2019)

[8] The deterministic laws of physics (e.g. classical mechanics) connect the complete state of the universe at a time with states at other times. I will return to this point later.

[9] The way we usually understand "S is up to me" it requires not only that my decisions influence whether or not S obtains but also that I know or have good reason to believe that my decisions influence whether not S obtains. Later I will discuss a version of the argument that appeals to this stronger understanding of "up to me."

states of affairs occur is for there to be dependence of the states of affairs on these decisions.[10] This dependence is expressed by subjunctive or counterfactual conditionals. For example, my decision at t to go to the university (decision to stay home) influences whether or not I am later at the university since if I were to decide to go (stay home) I would later be at the university."

A decision d influences a state of affairs o iff o (or the objective probability of o) counterfactually depends on o. That is, if both

d>o and -d>-o (d>P(o)=x and -d >P(o)=y) and x=/=y)

are true. In other words, which state of affairs (or its probability[11]) obtains counterfactually depends on which decision I make. Note that on this technical notion of influence it is not exactly the same as causation and a decision may influence a state of affairs even if the state of affairs occurs before the decision and is not caused by it.

The Consequence Argument formulated in terms of influence is

1) Determinism: The total state of the world at time t (and that it is the total state) and the laws imply the states of the world at all other times t'.

2) PAST: I have no influence over the past state at time t: there are no alternative decisions d1,d2 possible for me at t where s1,s2 such that d1>s1 and d2>s2 and where s1 and s2 are incompatible states of affairs that pertain to times prior to t

3) LAWS: I have no influence over the laws at t: There are no decisions d1, d2open to me at t such that d1>L1 and d2>L2 where L1 and L2 are incompatible laws

Therefore:

4) I have no influence over the future at t.

Since my having free will requires that my decisions influence the future it follows that

5) I lack free will.

Some philosophers reconcile themselves to the conclusion of the argument by accepting it and then arguing that lack of free will doesn't interfere with our practice of holding people responsible for their actions.[12] Even if one find this plausible those of us who think that we can influence and

[10] Van Inwagen's initial formulations of the argument and ensuing discussion obscured the role of counterfactuals in the argument. Lewis formulates a version of the argument in terms of counterfactuals and Kadri Vihvelin convincingly argues a proper formulation of the argument must involve counterfactuals. Vihvelin p 64-5 (2013) Van Inwagen in this response to Lewis accepts the counterfactual formulation.

[11] I will later discuss how probability should be understood in this context.

[12] This is the route taken by John Fischer in his *The Metaphysics of Free Will* OUP 194

thereby choose, at least to an extent, our futures the consequence argument presents a challenge to discover where it goes wrong.[13] There are only four possibilities i) the argument is invalid, ii) determinism is false, iii) Laws is false; one can influence the laws, iv) PAST is false; one can influence the past. I will discuss each in turn.

Rejecting the validity of the argument is not a plausible way out. The argument is intuitively valid and it is valid in standard Lewis-Stalnaker logics for counterfactuals.[14] Determinism may be false but since we don't know that it is false and we are interested in whether it conflicts with free will I will assume that it is true.[15] So that leaves LAWS and PAST (premises (iii) and (iv)). These strike many as obviously true. Laws are exceptionless and on most views possess some kind of necessity. How can we influence them? No one thinks that deciding to suspend the laws of gravitation while falling from a height will be of any avail. The past is fixed, over and done with. How can we have any influence over it? No one now thinks that she can now influence whether or not the milk spilled yesterday. Nevertheless, there have been attempts to show that one or the other of LAWS or PAST is false. I will discuss both.

Can we influence the laws? One thought which crops up in the literature on free will and laws is that on Humean accounts of laws there is a sense in which a person's decisions can influence laws. So, if Humeanism about of laws can been made plausible then this may seem to be a way to support rejecting LAWS. Helen Beebee and Al Mele call this response to the consequence argument, without endorsing it, "Humean Compatibilism." Here is how it is supposed to work. According to David Lewis' version of Humeanism, the Best Systems Account, laws are certain generalizations entailed by the axiom system that best systematizes the totality of fundamental events. Since my decisions and actions (and the more fundamental events that constitute them) are among the events that the best system systematizes there is a sense in which the laws depend on my decisions and actions.

According to Beebee and Mele the

> "…. Humean perspective, then, imposes a modus tollens on van Inwagen's quick argument quoted above–and, a fortiori, on the modal version of the consequence argument. Since, in his deterministic world, it is up to Fred whether or not he skips breakfast, and it is manifestly not up to Fred what went on in the distant past, it is up to Fred what the laws of nature are. We admit

[13] Another response is that the argument presents a mystery. There is something wrong but human beings are not capable of discovering what is wrong. This response is suggested by Peter van Inwagen in a recent talk.

[14] Both Lewis' (1973) and Stalnaker's (1968) semantics are based on similarity orderings of possible worlds have the truth conditions A>B is true iff there is a no A and -B world that is more similar to the actual world than and A&B world. Two differences are that for Stalnaker the ordering is simple and discrete while for Lewis there may be ties and the ordering may be dense. I will assume that the ordering allows for ties but is discrete. These differences make no difference to my discussion of the consequence argument. The argument is valid when formulated in Lewis or Stalnaker logics for counterfactuals. Suppose I actually decide d1 and o1 obtains. I have influence over o1 only if I am able to make a decision d2 such that d2>-o1. Unless I have influence over the past or the laws in the most similar worlds at which I decide d2 the past and the laws are the same as they actually are otherwise I would have influence over one or the other. But since the actual past and the laws imply the actual outcome if I have no influence over the past or the laws, I have no influence over whether it occurs.

[15] In any case, as I mentioned earlier there are other arguments lying in wait that claim to devour free will even if determinism is false Loewer 1996) Van Inwagen e.g. Van Inwagen's "roll back argument"

> that this may sound peculiar to some readers, but to Pat's Humean ears it is merely a way of saying that laws do not have the kind of ontological clout that would enable them genuinely to impede exercises of free will. According to the Humean view we have been assuming, there is no feature of the world that deprives Fred of the ability to skip breakfast tomorrow. "

Non Humean accounts of laws encourage the feeling that laws have ontological clout that constrain events including our decisions and actions. Jenann Ismael suggests that that this is partially what underlies the worry that determinism conflicts with free will.

> When we adopt a globalist perspective, our activities become part of the pattern of events that make up history. Since our activities partly determine the pattern, and the pattern determines the laws, our activities partly determine the laws. But then something weird happens. We invert the order of determination and reify the laws, so that now it looks like the laws are not simply descriptions of patterns that is partly constituted by our actions but are instead iron rails built into the spatial and temporal landscape that won't let us act in any way not in accord with them. (p. 111)[16]

But whatever the merits or defects of Humean accounts of laws Humean Compatibilism does not offer an escape from the consequence argument. The problem is while on a Humean account the laws do metaphysically depend on the actual events including my decisions and actions this kind of dependence is not the kind that is relevant to the Consequence Argument. That kind is that my decisions influence the laws, i.e. that the laws are counterfactually dependent on my decisions. But counterfactual dependence of the laws on my decisions doesn't follow merely from Humean account of laws. Humean accounts of laws are compatible with laws never counterfactually depending on my decisions. Whether or not decisions influence the laws depends on the truth conditions of counterfactuals not on the metaphysics of laws. Humeans can adopt the same recipe for evaluating counterfactuals that non-Humeans do. There are accounts of counterfactuals on which if L is a law and A and -A are compatible with the laws then both A>L and -A>L are true whether the laws are Humean or non Humean.[17] I will discuss such an account later.

Although David Lewis is the most prominent proponent of a Humean account of laws he does not rely on Humean compatibilism to respond to the Consequence Argument. He does reject LAWS but not on the basis of his Humeanism about laws but on the basis of his account of counterfactuals. Before discussing his response to the consequence argument, I will review his account of counterfactuals since it plays a central role in his response and problems with his account partly motivate my account and response to the consequence argument.

[16] Ismael's (2016) account of free will is subtle and full of insights about the nature of free will. I agree with her that reifying laws as certain non-Humean accounts do making them "enforces" of regularities seems to threaten free will, but I don't think that rejecting a non-Humean account of laws is sufficient to disarm the consequence argument. Ismael argues against both LAWS and PAST. While I agree that it is PAST that needs to go our reasons for rejecting it differ.

[17] Humean accounts of laws entail that whether or not a proposition is a law supervenes on the totality of non-nomological facts and so they will endorse counterfactuals of the form if A were the case then L would not be a law. But this is different from if A were the case then L would not be true, and it is this that is relevant to the Consequence Argument.

According to Lewis possible world semantics for counterfactuals A>B is true at @ iff B is true at all possible worlds most similar to @ at which A is true B is also true.[18] The truth conditions of counterfactuals depend on how "similarity" is understood. Kit Fine pointed out that if "similarity" is understood in an ordinary way then the counterfactual "if Nixon had pressed the button there would have been a nuclear war" would turn out to be false since a world in which there is no nuclear was is surely more similar to the actual world than is a world with a nuclear war. So, the relevant notion of similarity is not the ordinary one. What is it?

Lewis calls the type of counterfactuals related to influence and causation "non back trackers." Here is an example of a backtracking counterfactual: Suppose that I am on the 6th floor ledge and there is no safety net below me. Then the counterfactual "if I jumped, I would have landed safely since there would have been a safety net under me" understood as expressing a truth is a back tracker. But "If I had jumped, I would have been badly hurt when I hit the ground" is a non-back tracker. For non-backtrackers Lewis proposed an account of similarity that considers conformity to the actual laws as very important but allows for small local violations of the laws of @ at similar worlds. He calls these violations of @'s laws "small miracles." They are not literally miracles since they occur not in @ but in worlds similar to @ whose laws match @'s except for a small region. According to Lewis' well-known account the considerations involved in determining world similarity are:

1. Avoid big, widespread, diverse violations of law. ("big miracles")
2. Maximize the spatio-temporal region throughout which perfect match of particular fact prevails and maximize the time period over which the worlds match exactly in matters of fact
3. Avoid even small, localized, simple violations of law. ("little miracles")
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

Lewis' proposal is that the truth conditions for non-backtracking counterfactuals are given by semantics in which the similarity relation satisfies 1-4. These conditions are intended to have the result that the most similar worlds to the actual world at which a counterfactuals antecedent A is true match the actual world until the latest time at which a small localized violation of the laws leads to a state which then evolves by the laws so that A is true. For example, "If Nixon had decided to press the button at time t there would have been a nuclear war" is true according to Lewis since the most similar world W to the actual world @ at which Nixon decides to press the button is a world that exactly matches @ up until a short time before t when there is a very local violation of the actual deterministic laws   allowing for Nixon's decision to occur and then proceeds in conformity with the actual world laws to a nuclear war. So even if the state of the universe long before Nixon was born and the actual deterministic laws entail that he won't decide to press the button Nixon was able to press the button and his doing so would have influenced whether there is a nuclear war.

[18] More accurately A>B is true at @ iff there is a world W at which A&B are true and there is no world W* as or more similar than W to @ at which A&-B is true. This allows for there not being a most similar world to @.

Here are three quick observations about the truth conditions for counterfactuals based on Lewis' account of similarity that are relevant to my discussion of the consequence argument.

1) If determinism is true at @ then it is clear that the most similar world to@ which is even slightly different from @ at some time must be one in which either the history of the world is different from @ or the laws are different or both.[19] Assuming that counterfactuals are not trivial this is the case no matter how similarity is understood. On the account given by 1-4 if A is false at @ the most similar world to @ differs from it by including a small violation of @'s laws prior to the time t of A and has a history that also differs from the time of the violation to t.

2) Since the evaluation of typical counterfactuals with false antecedents like the Nixon counterfactual involves a world at which the laws of the actual world are violated in a small region one might suppose that this account goes along with a Humean account of laws. But this would be a mistake. The "miracle" occurs not in the actual world but in another possible world and it won't be a violation of that world's laws since laws of a world are never violated at that world. Both he laws in the actual world and the laws in the counterfactual world can be non-Humean governing laws. Nevertheless, a Humean account of laws seems more amenable to laws that apply except for a localized region.[20]

3) Lewis says that the notion of "similarity" characterized by conditions 1-4 a world in which Nixon presses the button but there is no war is less similar than one in which there is a war since big miracle is required to eradicate the traces of the small miracle and subsequent button pressing. This is essential to Lewis' claim that the truth conditions determined by 1-4 entail the temporal asymmetry of counterfactuals. What is meant by "temporal asymmetry" is that if E is a relatively local event that occurs at time t then if E had not occurred the past history prior to t would have been pretty much the same while the future after E might and often would be very different from the actual past and future. We will later see that Lewis is mistaken about this

Lewis uses his account of counterfactuals to respond to the Consequence Argument as follows: Assume that at the time at issue Nixon was able to decide either to press the button or not press the button. We think that had Nixon pressed there would have been a nuclear holocaust and Lewis' account seems, at first, to deliver this verdict. Whether or not there is a nuclear holocaust depends on his decision since it would take only a small violation of the actual laws in a counterfactual world prior to his pressing the button- say the firing of a few neurons in Nixon's brain- to lead by the actual laws to his deciding to press the button and thence to a nuclear holocaust. So, Nixon was able to influence whether or not there would be a nuclear holocaust. But Lewis' account also requires that whether or not there is a violation of actual laws counterfactually depends on his decision. In other words, if Nixon was able to decide either to press or not to press, he was also able to influence whether or not there would be a

[19] This point is driven home by Vihvelin

[20] A Lewisian best system for the counterfactual world in which Nixon decides to press the button might be one that is just like the best system of the actual world with the exception that in a small region just prior to t these laws don't hold but an event leading to the decision occurs.

small violation of law prior to his deciding. Lewis rejects both LAWS and PAST although his rejection is very qualified. LAWS holds except for a very small region where there is a violation of the actual laws and PAST holds except for the period of time between the violation and the decision.

Lewis' claim that we are able to influence the laws at first seems incredible. But he argues that it is simply a consequence of applying his account of counterfactuals to decisions. Further, once one recognizes exactly what that ability consists in it he think this can be seen as acceptable. It is the ability is to make a decision such that if we made it the actual laws would have been violated (though they would then not be laws) in a small region of space-time prior to the decision. He points out that possessing this ability doesn't mean that we are able to do something that is or causes a law-breaking event since the violation of law occurs prior to the decision. It doesn't give us the ability to suspend the laws of gravity or build a rocket that travels superluminally. Even though we have the ability to *decide* to make a rocket that travels superluminally we don't have the ability to make such a rocket. The ability to make alternative decisions doesn't violate the deterministic laws even though had we made a decision we didn't make laws would have been violated. Of course, we never do make the decision that requires an actual violation of law since that decision is made not in the actual world but in another similar world.

Recapitulating, Lewis claims that his account of counterfactuals captures the truth conditions of counterfactuals relevant to causation and influence. His response to the consequence argument is that it follows from his account of counterfactuals that LAWS is strictly false so there is no extra cost to rejecting LAWS to refute the Consequence argument. Van Inwagen agrees that this response succeeds in showing that the consequence argument is not a "slam-dunk" but he goes on to claim that the cost of responding to it as Lewis does is high since it involves accepting that our ability to freely choose requires "miracles."[21] Commenting on van Inwagen Kadri Vihvelin points out the cost of evaluating counterfactuals with false antecedents if determinism is true must be either violation of LAWS or violation of PAST so there is no additional cost for applying an account to decision counterfactuals.[22] Unless van Inwagen has an argument that shows that determinism precludes the truth of counterfactuals with false antecedents-which would be a truly remarkable result- Vihvelin's point is sufficient to show that the argument is unsound. Lewis' response shows one way how van Inwagen's argument might be unsound.

[21] He says that Lewis's paper is "the finest essay that has ever been written in defense of compatibilism – possibly the finest essay that has ever been written about any aspect of the free will problem". ("How to Think about the Problem of Free Will", *Journal of Ethics* (2008) 12, 337-341) But he adds that the Consequence argument "has nevertheless succeeded in raising the price" of compatibilism". (Freedom to Break the Laws", *Midwest Studies in Philosophy*, 28 (2004), Blackwell, 334-350).

[22] Vihvelin (2013) She argues that the consequence argument doesn't raise the price of compatibilism since assuming determinism any account of counterfactuals requires that the most similar worlds to the actual in which a false antecedent is true is one in which either the laws of the actual world or violated or the past of the antecedent differs from the actual past (or both). She claims that this observation is sufficient to disarm the consequence argument.

What does the work of disarming the consequence argument is not Lewis' particular account of counterfactuals but simply the observation that if determinism is true then possible world accounts of counterfactuals with false antecedents require either that laws are violated or that the past differs from the actual past. Lewis picks the first of these. While this is sufficient to defuse the consequence argument, I want to delve deeper into Lewis' response in terms of his account of counterfactuals since that account is defective. I will then sketch a different better account of counterfactuals on which PAST is false. This provides a different, and I will argue, better response to the consequence argument.

There are two problems with Lewis' account of counterfactuals: one relatively small and the other devastating. The relatively minor one is that it strikes many that the principle that if L is a law and A is logically compatible with L then A>L seems obviously correct.[23] It is because Lewis' account of counterfactuals violates this and thus allows for miracles that van Inwagen thinks that Lewis' response "raises the cost" of compatibilism. In fact, it follows from Lewis' account that if I actually make decision d1 then had I made d2 determinism would have been false. This does have an odd ring to it since it seems to say that even if both d1 and 2 are agent possible there is some sort of incompatibility between determinism and my ability to choose.

The devastating problem is that Lewis account of counterfactuals is not successful as an account of non-backtrackers. It endorses obviously false counterfactuals and it fails to capture the temporal asymmetry of counterfactuals and influence typical of non-backtrackers. Adam Elga showed that Lewis' account of similarity delivers the verdict that if Nixon had pressed the button there would have been no nuclear war and more generally that it doesn't account for the temporal asymmetry exhibited by typical back trackers.

Backtracking counterfactuals are temporally asymmetric in that local counterfactual antecedents that depart slightly from actuality can and often do lead to consequents that greatly depart from the actual course of events in the future but not in the past. For example, if Nixon had pressed the button at time t the course of events after t would have been radically different but the course of events prior to t would have at least been pretty much as it actually was. For example, there would still have been a cold war, Kennedy assassination and so on. Lewis believed that the truth conditions embodied in 1-4 captured this asymmetry. But Elga showed that due to the temporal symmetry of the fundamental dynamical laws this is wrong.[24] On conditions 1-4 a world W*at which Nixon presses the button and that satisfies the actual dynamical laws except for a small "miracle" after the button pressing is at least as similar to the actual world as the world W in which the miracle occurs prior to the button pressing. In W* the past of the button pressing is vastly different from the actual world's past while their futures match. In W it is the past that matches while the futures differ.[25] The consequence is that on Lewis

[23] Marc Lange even characterizes laws as propositions that satisfy the even stronger principle that if A is compatible with L and L is a law then A>L is a law. As mentioned in the previous footnote this is incompatible with Humean account of laws.

[24] Elga (2000) See also Loewer (2006)

[25] W* is constructed by temporally reversing the actual world from its end until a time shortly before the time of the button pressing when a "miracle" leads to the button being pressed after which W* continues in accord with

account since W* is as similar to @ as W is "If Nixon had pressed the button the past would have been pretty much as it actually was while the future vastly different" turns out to be false.

Elga's result is fatal to Lewis' account of counterfactuals and so fatal to his response to the consequence argument. One might try to repair Lewis' account. One way is to alter the account of similarity so that match with respect to the history prior to the time of the antecedent is more important than match after t.[26] This complicates the account and would mean giving up his hope of explaining time's arrows in terms of the asymmetry of counterfactuals. And we would still be left with the consequence that if we have the ability to freely influence the future, we also have the ability to freely falsify determinism. For these reasons I now want now to look at accounts of counterfactuals that respond to the consequence argument by taking the second horn of the dilemma. That horn is to reject premise PAST instead of LAWS. If determinism is true, then this means that the most similar world to the actual world will have a different past and future than the actual world. Accounts along these lines have mostly been neglected because it is thought that a small departure from actuality at t can lead by laws to very big differences in the past just as it does in the future. Lewis says

> "….there is no guarantee whatever that [a world where the actual laws are true and where Nixon presses the button can be chosen so that the differences diminish and eventually become negligible in the more and more remote past. Indeed, it is hard to imagine how two deterministic worlds anything like ours could possibly remain just a little bit different for very long. There are altogether too many opportunities for little differences to give rise to bigger differences. (Lewis 1979, 45)

Lewis' worry is that worlds which are macroscopically identical to the actual world at the time of the decision might if determinism is true have histories that greatly diverge from the actual history. In that case, it could turn out for example, that if Nixon had pressed the button Caesar might (or would not) have not crossed the Rubicon.[27] To avoid this he devised criteria which he thought resulted in the most similar worlds being one whose pasts are almost identical to the actual world. But he needn't worry. In classical mechanics (and in Quantum Mechanics) there are worlds that are macroscopically identical or almost identical to the actual world in the past of Nixon's counterfactual button pushing although they differ greatly to its the future. The account of influence counterfactuals that I will next

the actual laws. W* differs from W and the actual world in that its entropy increases in both temporal directions from the time of the button pressing.

[26] If one does this and also adopts NP, then the result is an account of counterfactuals very close to the account that Jonathan Bennett develops in *A Philosophical Guide to Conditionals.* Kadri Vihvelin employs it in her defense of compatibilism.

[27] Recently Cian Dorr in a recent paper laws are not violated as more similar to the actual world than worlds at which they are violated. Following Albert (2001) and Loewer (2007) he points out that doing this for worlds' with the kind of laws (classical mechanical or quantum mechanical) that we think hold at our world doesn't necessarily lead to worlds in which small difference from actuality lead to big difference in the past.

describe takes advantage of this. Instead of rejecting LAWS my account rejects PAST.[28] Setting up the account requires a detour into physics and the sketching of an account of counterfactuals based on this part of physics.

The part of physics that plays the central role in rescuing freewill from the Consequence Argument is statistical mechanics. Boltzmannian statistical mechanics was developed to explain thermodynamic regularities e.g. the gas laws, and in particular, the second law of thermodynamics. The second law says that the entropies of the universe and its suitably isolated subsystems increase over time until equilibrium (maximum entropy) is attained. The entropy of a system's microstate is measured by the size (relative to a standard measure) of the set of microstates that realize it. There are arguments (due to Boltzmann and others) that make it plausible that most (on the Lebesgue measure) of the microstates compatible with an isolated system not at equilibrium evolve in accordance with the fundamental dynamical laws to macro states with greater entropy. Boltzmann understood "most" as specifying a probability so argued that the second law should be understood as saying that the entropy of a system not yet at equilibrium very likely increases. An instance of this law is that a half-melted ice cube sitting in a tub of warm water evolves to a higher entropy state in which the ice cube is melted. Good! But Boltzmann soon realized that the because of the temporal symmetry of the fundamental dynamical laws it also follows that it is also very likely that the system evolved from a state of higher entropy i.e. that the ice cube was more melted prior to t. Bad! This result can be avoided by positing that the entropy of a system to which statistical mechanics is applied starts out low. Applied to the entire universe the proposal is that is entropy at the earliest time was very small. David Albert calls this assumption "the past hypothesis" (PH). He and I have proposed a fundamental theory that includes the PH and the probability distribution as laws that we call "the Mentaculus." The Mentaculus has three ingredients:

1. The fundamental dynamical laws
2. The past hypothesis
3. An objective uniform probability distribution over the micro histories compatible with the PH.

Albert argued that the Mentaculus entails not only that the entropies of the universe and its suitably isolated subsystems increase (or are very likely to increase) from earlier to later until equilibrium is attained but also that it plays a central role in grounding two other important temporal asymmetries- the asymmetry of records and the asymmetry of influence. Subsequently, I showed how the Mentaculus can be used to characterize counterfactuals involving decisions and argued that the PH. It is this account of counterfactuals that I will employ to respond to the consequence argument.

I won't repeat the arguments for Albert's and my claims here.[29] However, I do want to emphasize one point. Despite its name the PH does not presuppose the past- future distinction. Rather, it claims that at one time (around the time of the Big Bang) the entropy of the universe was very small. It doesn't

[28] A number of other responders to the consequence argument also suggest rejecting PAST. Among these are Dorr (2016), Esfeld (2019), Hoefer (202) and Ismael (2014). My account differs from these though it is closest to Dorr's.

[29] See Albert (2000) and (2015) Loewer (2007) (2020)

say that this time is in our past. If the Mentaculus succeeds in explaining the temporal asymmetries of influence, records, counterfactuals, and so on then it implies that this time is in our past and *explains* time's arrows. In this way the Past Hypothesis earns its name.[30]

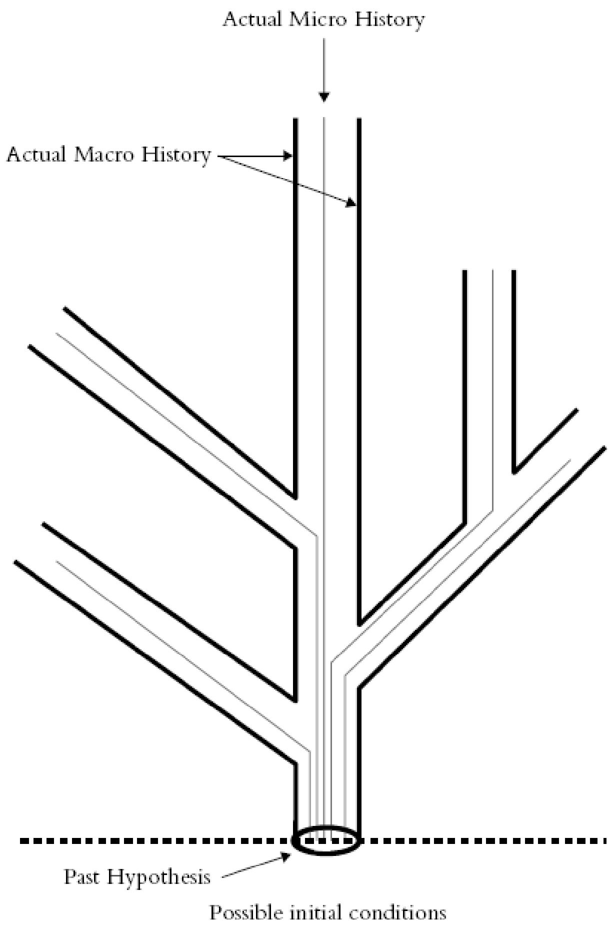
For the remainder of this paper I will show how the Mentaculus can be used to provide truth conditions for counterfactuals and how counterfactuals so characterized provide an effective rebuttal to the Consequence Argument.

The distinction between macro and microstates is central to understanding how the Mentaculus grounds temporal asymmetries. The microstate of the world at a time t is the complete fundamental physical state at t. In classical mechanics this consists of the positions and momenta at t of all the fundamental material particles. The thermodynamic macro state of a system S consists of the values of thermodynamic quantities- temperature, energy, pressure, average frequency of radiation, mass density etc.- in small volumes of S. More generally, the macroscopic state of the universe relative to a collection ∑ of macroscopic properties is the complete description in terms of the properties in ∑. A macro state is realizable by a continuous infinity of microstates. The entropy associated with a macrostate is the measure on the standard measure of the set of microstates that realize it. The fundamental claim of Boltzmann's statistical mechanics is that if S is a suitably isolated system whose macro state is M not at maximum entropy then the vast majority of microstates compatible with M evolve to macro states whose entropies are greater.[31]

The Mentaculus implies that while the evolution of the microstate of an isolated system is deterministic the evolution of its macro states is indeterministic where the probabilities of evolutions are the statistical mechanical probabilities. There are micro trajectories that are macroscopically identical until a time t when they macroscopically diverge. The result is a branching forks structure for the evolution of macro histories that looks like this.

[30] For elaborations and defenses of the claims see Albert (2000, 2012) and Loewer (2001, 2007, 2019).

[31] The reason is that in the neighborhood of every abnormal entropy deceasing microstate almost all (on the standard measure) states are normal entropy increasing state. It is also the case that every convex set of microstates contain some abnormal entropy decreasing states.



There are infinitely many micro-histories emanating from the PH that are macroscopically indiscernible until time t and then macroscopically diverge. There will also be microstates that are macroscopically distinct for a period and then converge. As long as the time is not too long after to the time of the PH divergences greatly predominate over convergences. The overall picture is one in which while determinism obtains on the microscopic level indeterminism reigns at the macroscopic level.

I will suppose, as seems plausible, that alternative decisions correspond to very small differences in brain states -say the firing of a few neurons- and that the macroscopic state at time t outside of the agent's skull is compatible with the agent choosing either d1 or d2. Two deterministic microhistories h1 and h2 may be macroscopically indiscernible at time t when they diverge with one realizing decision d1 and the other decision d2 a moment after t. Subsequent to t the macroscopic histories M1 realized by h1 and M2 realized by h2 may be very different. For example, M1 may include Nixon pressing the button and all that follows and M2 is the actual macro history in which he didn't press the button. The macroscopic state M at t doesn't determine what decision an agent will make a moment after t although the microscopic state that realizes M does determine the subsequent positions of Nixon's finger. This provides a sense of "possible" on which alternative decisions are possible for Nixon at t. Christian List's proposal that the ability to freely decide among distinct alternatives requires that alternative decisions and actions are possible is endorsed by the Mentaculus. List think that this is sufficient to defuse the consequence argument. But I don't see that this can be so since If I can't influence the laws and can't influence the past and determinism is true it will follow that I can't influence the position of my finger. To complete the reply, we need to show that either LAWS or PAST can be reasonably rejected. The Mentaculus shows why rejecting PAST is reasonable[32] To see how we first need to see how decision counterfactuals can be characterized in terms of Mentaculus conditional probabilities.

Decision counterfactuals express the probabilities of various outcomes if a particular decision were to be taken in current circumstances. For example, if I were now to decide to drive to the university the probability that I would arrive at the university in 20 minutes is x. I propose we evaluate decision counterfactuals as follows

$d1> P(A)=x$ is true if the iff the $P(A/d1\&M(t))=x$ where $M(t)$ is the macroscopic state a moment prior to d1

For example, if d1 is Nixon's pressing the button the probability of there being a nuclear war soon after d1 given the macroscopic state a moment prior to d1 will be very high. If d2 is his decision of not pressing the button, then the probability of there being a nuclear war soon after d2 is very low. So, Nixon has influence over whether there is a nuclear war. What about events prior to the decision? If E is a macroscopic event then except under very unusual circumstances $P(E/M(t)\&d1) = P(E/M(t)\&d2)$; E is probabilistically independent of d1,d2 given M. In other words, M screens off d1,d2 from E. And if $M(t)$ contains records of E then this probability is near 1.

Decision counterfactuals are temporally asymmetric. In this way the Mentaculus explains the temporal asymmetry of influence; why we can have some influence over the future but no (or almost no) influence over the past. "Almost no" because the account does entail that Nixon (and we) does have influence of the world's past micro history. Since determinism obtains the alternative decisions occur on different micro histories, m1,m2 of the world so the counterfactuals $d1>m1$ and $d2>m2$ are true. In other words, Nixon's decision influences the micro history of the world. But all this amounts that had Nixon decided to press the button the micro history of the world would have been different all the way back to the big bang. If Determinism and Laws are true and Nixon has influence over the future this has to be the case. But Nixon (and our) influence over the past is useless to him. He has no preferences for which past micro history realizes the past macro history. [33]

[32] See List (2018). List seems aware of this worry but tries to wriggle out of it by claiming that it is some kind of category mistake to combine agential and fundamental physical levels. But I don't see how this can show that the consequence argument is unsound.

[33] If there is not a record of an actual macroscopic event E, then the Mentaculus may assign a low probability of its occurring and so the relevant counterfactual d1 worlds are mostly ones in which E does not occur. But in this case whether or not d1 the probability of E is low, so the decision has no influence on E's occurring. This observation dispenses with Elga's "Atlantis" objection to the Mentaculus account of counterfactuals, See Dorr (2016).

So far my account applies only to decision counterfactuals. Showing how to extend it to counterfactuals in general is beyond the scope of this paper. But the basic idea is that in evaluating A>B where A is not a decision we look at the latest times at which there are micro histories that match the actual micro history macroscopically but the diverge so as to make A true and see how likely B is given A and the macro state at the time of divergence. The result is an account that superficially gives results that are much like the results Lewis intended his account to give. But it is also importantly different in a number of respects. The four most important differences are:

1) Unlike Lewis' account my account captures the temporal asymmetry of counterfactuals and is not subject to Elga's argument that devastates Lewis' account

2) Because on my account counterfactuals track conditional probabilities all macroscopic consequents of true counterfactuals with contingent macroscopic antecedents are probabilistic. For example, the account endorses "if Nixon had decided to press the button there *very likely would* have been a nuclear war' but not "if Nixon had decided to press the button there *would* have been a nuclear war." This is because there will always be some "maverick" microstates that are macroscopically identical to actual macro state which evolve aberrantly e.g. there is no subsequent war. But in the circumstances as we believe them to have been "most" of the microstates in the actual this macrostate evolve so there is a subsequent war. This is a virtue not a problem with the account. As Al Hajek has persuasively argued most counterfactuals with non-probabilistic consequents are false[34]

3) Instead of perfect match most of the micro histories in which the antecedent is true match only with respect to macroscopic events for which there are records. This has the consequence that the account violates PAST.

4) Instead of relevant alternative histories violating the actual laws they always conform to the fundamental dynamical laws until the time of divergence. So unlike Lewis' account LAWS is preserved."[35]

The Mentaculus response to the Consequence argument can now be clearly stated. It is to reject premise PAST by showing that given the physics of our world (i.e. the Mentaculus) and a proper understanding of influence freely chosen decisions that influence the future do also influence the past. But that influence over the past is negligible involving only microscopic events and useless since it doesn't enable us to exert any control over past events that are of any interest to us. Like Lewis' account my response has what may strike one as a counterintuitive consequence. Where Lewis' account implies that an agent has the ability to make a decision which if she had made it a law would have previously been violated my account implies that she has the ability to make a decision which if she had made it the microscopic past would have been different from the actual past. Where Lewis' account implies that if Newtonian dynamics correctly specifies the laws, I have the ability to falsify them my

[34] Hajek (2014). Hajek has a number of reasons for his claim one of which is similar to mine.

[35] In fact, not only is the Mentaculus not violated in these worlds but it continues to be a law even on Humean accounts of laws.

account doesn't have this consequence. Instead, it does say that if h is a specification of the actual micro history, I have the ability to falsify h though.

Both Lewis' and my account have a counterintuitive ring, but it is I think easier to come to appreciate that feeling that we can't influence the past is based more on mistaken views about influence and especially about time. It may be thought that one's having influence over a situation means that one can change it. But I cannot change either the past or the future though I can have influence over both. The difference is that my influence over the future may be significant while my influence over the past is negligible and uninteresting. Resistance to the claim that I have an ability to influence the past may be due to the picture of time flowing and that as it flows events are nailed down and once nailed down cannot be influenced. This is encouraged by thinking about time in terms of the so called "growing block". But past events are no more "nailed down" than are future events. Talk of "nailing down events", if it makes any sense at all, refers to the fact that we can influence future events in ways that we cannot influence past events.[36] Once it is realizes that all it means to say that we have the ability to influence the past is that the past would have been different if we made a different decision than we actually made the worry that we can't influence the past evaporates and with it the consequence argument.

The Mentaculus is a theory of the world that includes fundamental dynamical laws and laws specifying a low entropy boundary condition and a probability distribution over micro-histories compatible with it. This package accounts for the arrows of time- the second law of thermodynamics and the temporal asymmetries of records and influence counterfactuals. In this paper I applied the Mentaculus to show how free choice can be reconciled with physics. First, the Mentaculus yields indeterminism at the macro level and determinism at the micro level that accounts for the sense that the future is open so that an agent can chose among alternatives. Second, it shows how it can be that an agent can have significant influence over the future but not the past. This explains why our decisions appear to be undetermined and provides a sense in which we are able to choose among possible futures. More significantly it supports an account of influence counterfactuals that show how an agent's decisions can have significant influence over the future while having negligible influence over the past. This removes the threat that determinism and the consequence argument seemed to pose to freewill. If this is correct, then physics rather than undermining freedom it is part of its explanation.[37]

Albert, D (2000), *Time and Chance* Harvard University Press

Albert, D. (2015), *After Physics*, Harvard University Press, 2015

[36] See Carl Hoefer's (2002) for an insightful discussion of how mistaken views about time underlie the alleged conflict between free will and determinism.

[37] Of course, physics is only a part of the explanation of free will since at most it makes free will physically possible. Free will involves making decisions on the basis of reasons and rational deliberation. Accounts of these are the business of psychology and rationality theory.

Dorr, C., "Against Counterfactual Miracles" *Philosophical Review* 125 (2016): 241–86

Elga, A. (2000). 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence,' Philosophy of Science suppl. vol 68: 313–24.

Esfeld, M. (20i9) "Super-Humeanism and free will", *Synthese*

Hoefer, C. (2002) "Freedom from the Inside Out", in Callender, C. (ed.), *Time, Reality and Experience*. Cambridge University Press.

Ismael, J. (2016) *How Physics Makes Us Free* Oxford University Press

Lewis, D. (1973) Counterfactuals; Blackwell & Harvard U.P.

Lewis, D, (1979), "Counterfactual Dependence and Time's Arrow," *Noûs*, 13: 455–476.

Lewis, D. (1986) *Philosophical Papers* Volume II (1986; Oxford U.P.)

List, C. (2019) *Why Freewill is Real* Harvard University Press

Loewer, B. (2007) "Counterfactuals and the Second Law" in *Russell's Republic* ed Corry and Price

Loewer, B. (2020) "The Mentaculus Vision" *in Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature,* World Scientific (May 2020).

Stalnaker, R. (1968) "A Theory of Conditionals" in *Ifs*

Van Inwagen, P. (1975) 'The Incompatibility of Free Will and Determinism', Philosophical Studies 27: 185-199.

Van Inwagen, P. (1983) An Essay on Free Will, Oxford (Clarendon Press).

Van Inwagen, P. (1989) 'When is the Will Free?', Philosophical Perspectives 3: 399-422

Van Inwagen, P (2000) "Free Will Remains a Mystery" - *Philosophical Perspectives* 14:1-20

Vihvelin, K. (2013) *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*

Van Inwagen doesn't object to Lewis's way of stating his argument. On the contrary, he has said that Lewis's paper is "the finest essay that has ever been written in defense of compatibilism – possibly the finest essay that has ever been written about any aspect of the free will problem”. ("How to Think about the Problem of Free Will”, *Journal of Ethics* (2008) 12, 337-341).

Van Inwagen now agrees that the Consequence argument fails as a reductio.

However, he claims that it has nevertheless succeeded in” raising the price" of compatibilism. (Freedom to Break the Laws", *Midwest Studies in Philosophy*, 28 (2004), Blackwell, 334-350).