24.221 Final Paper

May 13, 2015

Determinism states the following: given the state of the universe at time t_0, denoted S_0, and the conjunction of the laws of nature, L, the state of the universe S at time t is uniquely determined. When I say, "the state of the universe", I mean some description that captures all of truths about the universe at the time in question. This includes information such as: "is my hand raised?", "is it raining?", and "what's the velocity of a particular molecule of oxygen?". In a famous argument by Peter van Inwagen, he concludes that determinism is incompatible with free will. This conclusion follows from a proof that he does to show that what we can do and what we do coincide (in some sense of "can do"). This deeply troubling conclusion has created a lot of rebuttal by people rightly titled compatibilists. There are many arguments for compatibilism, which I will not describe here. In this paper, I will set up van Inwagen's famous argument, and provide an original argument for compatibilism. I will argue that due to our vast uncertainty about the future, we experience thoughts which causally influence our actions in a way that captures our intuitive notion of free will. Therefore, despite the fact that van Inwagen's proof is valid, it doesn't challenge what we care about when we think about free will.

When van Inwagen talks about how he wants free will to be understood, he says that "it seems to be generally agreed that the concept of free will should be understood in terms of the power or ability to do otherwise. To deny that agents have free will is to assert that what a man does do and what he can do coincide." So for van Inwagen, a necessary condition for free will is the "ability to do otherwise," something that I'll later contest is not a necessary condition (which will require on expanding the meaning of "ability to do otherwise"). To prove the incompatibility between free will and determinism, van

Inwagen presents a case in which there is a judge, which, after careful deliberation, decides not to raise his hand. As a result of which, he does not raise his hand. We'll call the not-hand-raising action H. Despite the fact that we would normally think of this as a free action, van Inwagen then shows that the not-hand-raising is the only thing that the judge could do, and therefore, what he did do, and what he could do coincide, meaning that the judge does not have free will. The argument is very simple: H is entailed by S_0 and L by determinism. Now, take the phrase "the judge could have done otherwise" to mean that he could have made $\neg H$ true. van Inwagen uses the equivalent claim that the judge "could have rendered H false". But because $S_0 \wedge L \rightarrow H$, it is also true that $\neg H \rightarrow \neg(S_0 \wedge L)$. But this means that the judge could render either the initial state of the universe false or the laws of physics. The judge can't do that. This follows from two basic beliefs: one being that the character of natural law is such that no one can change them, and the other is that we can't edit the past[1]. Many compatibilists and incompatibilists would believe this statement without further elaboration on what it means[2]. Since the judge can't render the conjunction of S_0 and L false, he couldn't have rendered H false, i.e; he had to raise his hand. It was the only thing that we could do. Obviously, this argument generalizes to all sorts of actions, and therefore, we could not do otherwise. Free will is incompatible with determinism.

Before arguing that our notions of free will are compatible with determinism, I want to point out that while van Inwagen's argument seems simple and the conclusion unavoidable, it's a conclusion that many philosophers (and regular people) find deeply troubling for many reasons. I'll briefly describe two. One is that if our actions are already predetermined, why should we bother thinking about what to do with our futures? Whether or not I think about it won't change whether or not it happens, right? The second is also of pragmatic concern: we hold beliefs that people are responsible for their actions if they had the ability to do otherwise. If determinism is incompatible with the ability to do

[1] Even if we allowed for time travel, this isn't possible. Lewis argues in his paper on paradoxes in time travel that the past is uneditable. This is used to avoid grandfather paradoxes.

[2] I should point out that Lewis does not think the statement "we can't violate natural law" is specific enough. He thinks that there is a distinction between "weak abilities" to violate laws, and "strong abilities". I will ignore this for now, and assume that the claim that we can't violate laws is specific enough

otherwise, then it appears that no one is responsible for their actions. If that's so, then how do we punish people for wrongdoings? The reason that I point these two problems out is as follows: we have a lot of empirical evidence/beliefs to think that thinking about our futures is worthwhile. We also have a lot of evidence to believe that there are some instances where someone is clearly responsible for their action, while there are other instances where it seems like they're not. For example, if someone is coerced or forcibly consumes some judgment-impairing drug and then kills someone, we're less likely to see the murder as their fault. These two experiential claims give us some reason to be suspicious about the force of van Inwagen's conclusion, even if it is valid. Furthermore, both of these dilemmas directly concern the "ability to do otherwise". This gives us reason to think that the way out of van Inwagen's conclusion is to reject that free will has to do with the ability to do otherwise. What I will now do is argue that free will has to do with an "ability to do otherwise" that is different from van Inwagen's "ability to do otherwise". Then I will give an account for how introducing indeterminacy in our own futures lets us change our future actions in a way consistent with determinism. Then I will re-define free will and argue that this definition has the right features to preserve our pre-philosophical intuitions about free will.

van Inwagen comes to two conclusions in his paper. The one that directly follows from his premises is that the judge can't do anything but not raise his hand. The second is that free will and determinism are incompatible. The first follows from well-defined premises. The second is a generalization of the conclusion of the first argument. One way we can reject van Inwagen's second conclusion without rejecting his first conclusion is by making a separation between "historical possibilities" and "physical possibilities". Historical possibilities will be taken to refer to those things which are consistent with S_0 in our world and L. Physical possibilities can be taken to refer to things that are consistent with L, but not necessarily S_0. Physical possibilities can also be taken to refer to things consistent with S_0, but not necessarily L, but only if the inconsistency with the laws follows from our ignorance about them[3]. In van Inwagen's argument about the

[3]This is a delicate point for a couple of reasons. The main reason, in my opinion, is that this promotes physical possibilities to a subject-dependent notion. You may think something is physically possible, while I may disagree. Example:

judge, the judge raising his hand is not a historical possibility, because it's inconsistent with S_0, L and determinism. On the other hand, the judge raising his hand is certainly physically possible for most of us, for example, if we alter S_0, i.e; if the initial conditions were different. Even leaving S_0 fixed, we don't know nearly enough about how the mind works to conclude that the laws necessarily don't allow for him to raise his hand in the future. Therefore, van Inwagen's conclusion might read: "what the judge can historically do is always the same as what he does." This suggests that we might rescue free will by rejecting van Inwagen's criteria for free will such that it is associated with the "historical ability to do otherwise".

The next thing we have to do with this new criterion is show that it's consistent with our intuitive notions of free will. I think this is a necessary thing because the most disturbing feature of van Inwagen's argument for many people is that it challenges our deeply ingrained belief that we have free will. What are, intuitively, the important characteristics of free will? Obviously there are many, but I'll focus on the following:

1. The capacity to correct ourselves. A good notion of "we can do otherwise" should explain the observation that when we do things that we regret doing, we try, and often succeed in correcting them in the future. For example, the judge not raising his hand meant that someone was executed. Suppose that a similar trial happens in the future and the judge is once again put in a position where if he doesn't raise his hand, this new criminal will also be put to death. Suppose that between the first and second trials, the judge had a change of heart and felt that death sentences are bad and that he would no longer sentence people to death[4]. In future trials, we observe that the judge never sentences anyone to death again. This scenario is representative of something that could easily happen in real life. I'll refer to this item

you regret a decision that you make and you say "if I have a specific neuronal firing pattern F in the future, I won't make that decision again, and it doesn't seem like I couldn't do F, so F is physically possible. Me, as a well-trained neuroscientist, know that such a firing pattern isn't possible, and that you're doomed to make this decision again. Therefore, if you ask me whether F will be a physical possibility, I'll say no. I think that this is okay, because often we are mistaken about what we can and can't do, yet this mistaken belief that we can do otherwise also seems to be a part of free will because it will undoubtedly influence other actions that we do in the right way to be considered free. I discuss a criterion for an action to be free later in the paper, so this last claim about influencing free actions will become more clear then.

[4]I'm using the language of feeling versus deciding because an incompatibilist might argue that we don't decide anything due to determinism because we're not freely acting agents. But certainly determinism isn't inconsistent with the neuronal firing patterns associated with feeling.

as item 1 from now on. I also would like to focus centrally on this problem of how we self-correct. I think it is important to understand self-correction, because a priori, it seems like determinism challenges this. It seems to say that we're constrained to act a certain way that was pre-determined from the time that the universe started. So how can we correct ourselves?

A central assumption I make is that if we can give an account of self-correction that is consistent with determinism, then we rescue the sense of "we can do otherwise" that is really important to us. In particular, since I think self-correction is essential to free will, if we can show self-correction to be compatible with determinism, the language of free will isn't too relevant anymore. It's a proxy term for the capacity to self-correct and shape ourselves the way we want to be.

2. The second thing, which I won't focus on as much, but is important, is the ability to take responsibility for our own actions. If we give a proper account of this, then we can circumvent the second problem and claim that having a justice system is justified.

Right now, I will give an account of self-correction consistent with determinism - in line with the program stated at the end of item 1. I'm going to show why physical possibilities are relevant in this account, rather than historical possibilities. In doing so, I want to create an association between physical possibilities, and self-correction. And using the assumed connection between self-correction and free will, I will thus establish a connection between physical possibilities and free will, thus directly countering van Inwagen. To understand the role of physical possibilities, we need to invoke the fact that nobody knows the full state of the universe at any earlier time, S_0, or all of the laws. The state of the universe involves the specification of a vast number of degrees of freedom. Therefore, when we try and evaluate future outcomes, there is a correspondingly vast uncertainty in what we think can happen. Of course, by determinism, only one of these possibilities will be actual (historically consistent), but we have no idea which one that will be. It is experimentally true that we're always at least a little uncertain about the future.

This uncertainty stems from the vast number physically possible futures. Therefore, when we think about the future, our reasoning is influenced by physical possibilities, and not historical possibilities. It doesn't make sense to reason in terms of the uniquely determined future, because no one has any idea what it is.

For the judge, knowing that it's physically possible for him to not sentence people to death (shortened to "not kill people" from now on) in the future is going be associated with neuronal firing patterns (a blanket term that I'm defining to mean "whatever mechanisms are responsible for a particular thought or action") that say "I can change the way that I act and not kill people". I take this claim to be an empirical statement about psychological regularities. Combining this with the intuitive claim that these neuronal firing patterns are correlated with future actions that reflect the judge's decision to not kill anymore, we can establish a correlative connection between the judge's corrective action (namely, no more killing), and the physical possibility of not killing. Generalizing this example with the judge, I can claim: the fact that we reason about the future in terms of physical possibilities rather than historical possibilities is reflected by neuronal firing patterns that are typically followed by corrective actions. Furthermore, we often think about our futures because they, due to our limited knowledge, are indeterminate. This thinking leads to neuronal firing patterns that make us act in a way consistent with our optimal vision of the future (in typical people). Then thinking about the future is trivially beneficial. It's also a nearly inevitable consequence of indeterminacy that we can think of as following from psychological laws. Now, we have an account of self-correction completely consistent with determinism.

It's worth repeating the conclusion this account using different words in order to make it more clear what's going on: despite the fact that our future actions are uniquely determined by determinism, determinism does not remove the possibility that these actions are influenced by neuronal firing patterns in a causal way. Determinism also does not remove the possibility that after we do something, we experience thoughts that lead to actions that cause us to not do that thing again (or to keep doing it). These thoughts we experience come from a (not necessarily) correct belief that we can do otherwise,

consistent with our extreme uncertainty about the future state of the universe. [5]. This account has the nice feature that it allows for events to causally determine our thoughts, which causally determine our future actions. In other words, our actions flowing from our desires or thoughts. This is important because in van Inwagen's setup, we had the judge, which *after careful deliberation*, did not raise his hand. We intuitively thought of that as being a free action on his part because his action followed directly from his thoughts or desires. Therefore, my account rescues our intuitive requirements for a free action. In what follows, I will address some possible holes left unfilled by my account, and use my account to motivate a definition for a free action.

One seemingly serious objection to my account of what matters in free will is that I don't address whether or not we are the agents responsible for the actions that we commit. In particular, although the judge experiences these thoughts of regret about killing people, and his underdetermined knowledge about the future leads to thoughts consistent with his not killing people in the future, it is not obvious that there's a sense in which the we can say the judge "chose" to have these regrets in the first place. The neuron firing patterns that I describe in the last paragraph were inevitable by van Inwagen's argument. Maybe this means that it doesn't quite make sense to attribute the judge's action to him. In which case, maybe it feels like we still don't have free will. Our neuronal patterns, and thus our beliefs, values, choices, and consequent actions are as determined as the motions of a puppet that's being controlled by a puppetmaster. Before I respond, I'd like to note that the point of the objection isn't to save van Inwagen's second conclusion about us not having free will, but rather to enhance my own account of what matters in free will, in order to further convince us that despite determinism, we have everything we want with regards to acting freely. This of course has the side benefit of diminishing the force of van Inwagen's second conclusion, but only indirectly by making my account more complete.

My proposed response is a rhetorical question: does it really matter that it doesn't

[5]This non-trivial psychological "fact" is crucial to the argument, and it's worth briefly motivating. Suppose we worked on the belief that we couldn't do otherwise (i.e; by reasoning about historical possibilities). Then it would not be worth our time to think about how to do otherwise. For example, I believe (correctly) that it is not possible to accelerate a particle past the speed of light, so why bother thinking about how to do it?

seem like we can attribute desires and actions to ourselves? Let's *define* an action as being a result of *our* free will, and therefore due to us, if it directly follows from our desires. In other words, if my raising my hand follows from the neuronal firing patterns that tell me that I want to raise my hand, and these neuron signals make me raise my hand, then I raised my hand by my own will. A not-free action is one such that, despite my will telling me that I don't want to do it, I did it anyway. This includes things like pulling a trigger due to a nervous twitch, despite not wanting to. It doesn't include cases where I marginally decide one way or the other. For example, if I'm deciding whether or not to shoot the aforementioned gun, and it's a hard decision, and I eventually decide to shoot it, this is a manifestation of my will. It was a free action.

This is a very objectionable definition. In particular, someone who disagrees with me will immediately say that I can't "define" free will like this because it avoids the original objection by re-defining free will to make the objection invalid (basically equivocating). Originally the question that I needed to respond to was whether or not we can attribute to ourselves the actions that we do. If I just define them as free, the question is trivialized. But the question doesn't appear to be trivial, which is suggestive of an equivocation. I think that in order to resolve this problem, I have to show that our original worry, or alternatively empty. By empty, I mean that we could equivalently say that "we don't attribute actions to ourselves" or "these actions belong to us". The form of equivalence I have in mind is this: whatever answer you choose, you can make all of the same predictions. One way to test for equivalence of this form is to ask if you can perform some kind of conceivable physical experiment to distinguish between the two answers. I believe that the question of "who performs the actions" is empty. In part, this is because "the initial state of the universe and the laws of physics" are not objects. They're propositions that make statements and/or predictions about the objects in the universe. And we can only observe events which are carried out by objects. So if I see someone do something, how could I possibly preferentially attribute it to "the laws of physics and the initial state of the universe", or nothing at all, rather than the person doing the action?

One possible counter to the emptiness claim is this: "if the question of 'who did

it?' was empty, then I shouldn't care what I attribute the action to because everything observable is the same. But I certainly do care about what to blame! Think about the justice system, where we need to identify the entity that does a bad action so that we can penalize them accordingly. If I can't identify the entity performing the action, then how can we definitely hold anyone responsible for wrongdoing?" My response is this: in a situation like a justice system, the practical goal is to keep people safe from wrongdoers, or to preserve order in society. It at the very least has to achieve some physically observable goal. We should make the choice of holding people responsible for their actions. If we don't, then we don't penalize anyone and then don't achieve whatever goals we have for our justice system on purely practical grounds. If we take that the question of attribution is empty, we can at least say that it isn't *logically wrong* to hold people responsible for their actions. Normal people seem to have the ability to correct their actions given their indeterminate knowledge of the future and influence their later actions, i.e; to execute the account I gave for self-correction. At least, this is a ability that we ascribe to normally functioning people. For people who are incapable of self-correction, the question becomes more complicated, and out of the scope of the original problem we were addressing.

To summarize what I have done here, in a slightly more illuminating order: we started with the apparent problem posed by van Inwagen that our inability to do otherwise leads to our not having free will. This problem rests on his crucial assumption that free will has to do with the ability to *legally/historically* do otherwise. I showed this by showing that it's possible for the judge to do otherwise if we're talking about physical possibilities. I claimed that the free will that we intuitively know is related to physical possibility and not historical possibility. To address free will, I defined an action as following from our free will it follows from the desires encoded in my neuronal firing patterns. This definition is not incompatible with determinism since it's a criterion in terms of consistency between neuronal firing patterns and observed actions if we ignore the possibility of external influences like coercion. And determinism certainly doesn't invalidate this consistency. In fact, this consistency just follows from physical law. To show that we can recover our pre-philosophical intuitions about free will, which I claimed was very important because

these intuitions were precisely what incompatibilism was challenging, I gave an account of actions under this picture of physical possibilities mattering. Namely, the fundamental indeterminacy in the future due to our incomplete knowledge of S_0 leads to the following phenomenon: I think it's physically possible to do something, and if I want to do it, I experience neuronal firing patterns that are consistent with wanting to do it, and they force actions that try to do that thing. This allows me to have the ability to correct actions that I think are wrong and not do them again in the future. It makes sense to call this process "acting with free will" because it captures something fundamental about free will: that we can do things that are consistent with our desires. It doesn't matter that it's the only thing that could have historically happened. Our intuitive notions of free will don't reflect that.

24.221 Metaphysics
Spring 2015