"Philosophy is the strangest of subjects: it aims at rigour and yet is unable to establish any results; it attempts to deal with the most profound questions and yet constantly finds itself preoccupied with the trivialities of language; and it claims to be of great relevance to rational enquiry and the conduct of our life and yet is almost completely ignored. But perhaps what is strangest of all is the passion and intensity with which it is pursued by those who have fallen in its grip."

♦

Kit Fine

CHIEF EDITOR

Mason Westfall

PRODUCTION EDITORS

Selin Akkan
Nancy Guerrero
Andrew Liedholm
Mason Westfall

FACULTY REVIEWERS

Professor Rick Grush
Professor Saba Bazargan
Professor Dana Nelkin

STUDENT REVIEWERS

Selin Akkan
Andrew Liedholm

A NOTE FROM THE EDITORS

Every year hundreds of papers are written for philosophy classes at UCSD. These papers cover an extremely broad range of material, both historically and topically. Students spend many late nights struggling with difficult texts and wrestling with subtle arguments. That hard work produces a heterogeneous array of arguments and interpretations. It is the aim of this journal to present the cream of that bountiful crop, to showcase the finest achievements in undergraduate scholarship of the past year. These students created notable works in Philosophy of Action, Ancient Philosophy, Philosophy of Science, Decision Theory, and Political Philosophy. We are proud to be able to share these papers with a broader audience. Please enjoy *Intuitions*.

ACKNOWLEDGEMENTS

COVER ART:

The Treachery of Images, Rene Magritte

CONTENTS

AGENT CAUSATION AND FREE WILL

— Eric Della Santa —

The Dilemma Argument makes a good case that, whether the world is deterministic or indeterministic, there is not much room for free will of the person to influence events. Its general form goes something like the following[*]:

1) Our actions are either determined or they are not.

2) If they are determined, then we cannot do otherwise.

3) If we cannot do otherwise, then our actions are not free.

4) Therefore, if our actions are determined, they are not free.

5) If our actions are not determined, then they are a matter of luck.

6) If our actions are a matter of luck, then our actions are not free.

7) Therefore, if our actions are not determined, they are not free.

8) Therefore, our actions are not free.

The reason this argument is so compelling is largely because the manner in which we determine which action to perform is generally considered to be an event in itself, therefore must have been causally influenced by prior events. The way some libertarians have found to get around this problem is by introducing the agent causal view, wherein agent causation is a whole other kind of causation alongside event causation and it doesn't have to be causally determined by any prior events. That is a broad picture of the agent causal view at least, and there are many variations of it, as is true of any view. One such view that is both prominent and insightful is the view presented by Timothy O'Connor[†]. In paper I will present and critique O'Connor's view as an example of a strong agent causal view and where it still fails, and then I will present my own view on how agent causation might overcome its difficulties.

To begin his argument, O'Connor defines exactly what he means by the term 'causation'. Although this matter may sound trivial, it is actually a very important foundation because, depending entirely on how this term is defined, agent causal theories can either be a possible candidate to fit the mold, or be rendered completely

---

[*] I am indebted to Prof. Dana Nelkin for this summary of the Dilemma Argument, as well as for providing me with many other useful insights on this topic.

[†] O'Connor, Timothy. "Agent Causation." *Free Will*. Ed. Gary Watson. Oxford: Oxford University Press, 2003. Print.

incoherent. O'Connor defines causation as a primitive notion of the "'bringing about' of an effect" that defies the ability to conform entirely to any real causation theory. With this definition he does not inherently include or exclude the ability of agents to produce causation. However, he does say that objects have causal powers. This is why events can cause other events. Under his theory, agents, too, have causal powers, though they are more special than event causal powers. Agents have the ability to directly determine their action in more than one direction. So while any given event can only produce one specific effect, agents can produce multiple possible outcomes. This also bases his theory in indeterminism, because if the world were deterministic then there would be no room for agents to produce events of their own without being a prior event cause themselves.

The other ability agents have, according to O'Connor, is the power to cause and arrange intentions. Thus, when an agent performs an action, he or she also causes an intention for that action. Furthermore, the agent can determine which intention to cause for any given action among the various intentions available for that action. This allows O'Connor to factor in reasons an agent might have for performing an action and to say which reasons it is that the agent acted on. Other attempts to factor in reasons would include Randolph Clarke's agent causal view[*] where reasons are causes alongside agent causation, or really any compatibilist account of free will which considers reasons to be the sole causes themselves. A problem with any solely agent causal view is that we want our actions to be completely open and undetermined, but we also want our reasons for doing things to play a role in our decision. But since a solely agent causal view cannot allow for reasons to play a causal role, differentiating between the reasons we have for doing something and the reasons we actually acted on can be difficult. I will discuss this more when considering objections to his view, but for now it is important to note that by actually finding a non-causal role for reasons in the intentions we cause, O'Connor made a valiant effort at overcoming this difficulty, though his success at doing so is questionable.

One major objection to O'Connor's view is an objection common to all libertarian accounts of free will. Since he does not allow for reasons to play a causal role, he cannot give a contrastive explanation to account for why an agent may choose one action for one set of reasons over another action for a completely separate set of reasons. For example, an agent may be craving both a hot dog and a hamburger and is at a fast food joint that serves both. A view that allows for reasons to play a causal role can say that the agent chose, say, to order a hamburger instead of a hot dog because the reasons of it being cheaper and it more filling were more powerful reasons than the hot dog

---

[*] Clarke, Randolph. "Toward a Credible Agent-Causal Account of Free Will." *Free Will*. Ed. Gary Watson. Oxford: Oxford University Press, 2003. Print.

being more delicious. However, O'Connor's view can only say that the agent made his or her decision to order a hamburger for the reasons that it was cheaper and more filling. But it cannot say why he or she did this action for these reasons instead of ordering a hot dog for the reason of it being more delicious. The second element to this objection is that O'Connor's view cannot offer the kind of contrastive explanation that accounts for why an agent performed an action *when* that agent performed the action. Once again, normal event causal theories can offer this explanation easily. But this kind of contrastive explanation cannot be given (though some, such as Clarke, have attempted to provide one) through agent causal theories because the agent, not being an event at a given point in time, exists outside of time, in a manner.

O'Connor answers these objections by agreeing that his view truly cannot offer either kind of the contrastive explanation discussed. However, he also asserts that this is not a problem. In fact, it is not even a phenomenon unique to agents, considering the fact that quantum mechanics runs into the same problem with completely randomly occurring events. Therefore, he argues, we should not be surprised that agents cannot offer a contrastive explanation, nor should we object to the lack of it because it cannot detract from the coherence of the theory. As long as his view can account for reasons, on some level, in the agent's decision, which it can, then that is all the explanation we should require.

A second objection to O'Connor's view concerns that fact that reasons do not play a causal role under his account of agent causation. As he notes, this objection is raised from a now widely held belief propagated by Donald Davidson called the "reasons are causes thesis"[*]. According to this thesis, the reasons that explain actions are reasons that cause action. This thesis gives us the ability to distinguish between reasons we act on and reasons we have but don't act on. Many have now taken it as the only way to form such a distinction, so when O'Connor explicitly says that reasons are not causes under his view, it raises the question of how he can distinguish between reasons an agent acts on and reasons that exist but the agent doesn't act on. As I mentioned earlier, O'Connor attempts to get around this objection by finding a role for reasons, even though it is non-causal. He argues that the reasons we act on are referred to in the intentions the agent creates when he or she performs an action. In this manner he differentiates those reasons from other reasons the agent might have to perform the action that were not actually acted upon. Although his method of doing so can be fairly unclear and confusing, the mere fact that

_____

[*] Davidson, Donald. "Actions, Reasons and Causes." *Journal of Philosophy*. 60.No. 23 (1963): 685-700. Print.

O'Connor found a way to make that important distinction without invoking Davidson's theory is worthy of note because it shows that the reasons are causes thesis is not necessarily true even if the rest of his theory is rejected.

O'Connor's theory thus holds several key advantages. It offers the capacity for free will that we have been looking for without having to show how it could be compatible with deterministic or indeterministic event causation. Furthermore, it has achieved more than most other agent causal theories by establishing the coherence of his account and allowing for reasons to play a role while not falling into the trap of making it a causal role. This allows him to advance a more pure agent causal view than competing views such as Clarke's. However, it also has some problems that O'Connor does not specifically address. For example, while he did establish that his view can be coherent without contrastive explanations, that does not necessarily increase the amount of control we have. He would say that as long as the agent is the direct source of the action, we have all the control we need, but it is a legitimate question whether we could have the kind of control that most of us desire without the ability to offer a contrastive explanation. After all, how can we say that we have full control over our actions if we cannot even explain why we favored one set of reasons over another? Another argument against O'Connor is, of course, Galen Strawson's self-creation argument[*]. Nearly any agent causal view is really going to fall prey to this argument because they would need to show how they can be responsible for certain traits of the agent, and this would end up requiring an incomprehensible infinite regress. Finally, truly skeptical arguments such as the one given by Derek Pereboom[†] cannot be answered by O'Connor's view. To answer them, he would have to explain how agent causation can be physically possible, considering the fact that the neuroscience we have today points toward the conclusion that the brain simply doesn't work in a manner that would allow for an agent causal view[‡]. Therefore, despite the strength of his view in general, skeptics like Pereboom will remain unconvinced. However, it is also true that none of these unanswered objections can really blow a hole in O'Connor's argument to the point that it would be indefensible

---

[*] Strawson, Galen. "The Impossibility of Moral Responsibility." *Free Will*. Ed. Gary Watson. Oxford: Oxford University Press, 2003. Print.
His argument, which he calls the Basic Argument, is that to be truly morally responsible, you must be responsible for how you are in certain mental respects. But to have this kind of responsibility requires an infinite regress that is impossible to achieve. Thus it is called the self-creation argument because it implies that we would need to essentially be the creators of ourselves, an idea which he agrees with Nietzsche on in that it is absurd.

[†] Fischer, John, Robert Kane, Derek Pereboom, and Manuel Vargas. *Four Views on Free Will*. Malden, MA: Blackwell Publishing, 2007. 85-125. Print.

[‡] Pereboom adopted a position of hard incompatibilism. This meant that he did not think that we could have free will regardless of whether determinism or indeterminism is true because in neither occasion would we have the kind of control necessary for free will. He did think agent causal views were the exception to this, though he believes science has shown that that just isn't the way the world works.

(except for maybe Strawson's). The most they can do is offer reasons for not being convinced by it. And so, in that respect, O'Connor has given us a semi-successful account of agent causation.

As an agent causal libertarian myself, I think O'Connor has thus far presented the most satisfying account of free will. However, I also take issue with many of its nuances. My biggest objection is to the way he incorporates reasons. Causing an intention for an action at the same time as performing that action is as bewildering of an idea as it is unsatisfying. I do not know how it is that we can cause intentions, or even what form those intentions might take if not the causes for our actions. Furthermore, I do not think we want to simply do an action without reason having some kind of direct impact on our choice, and then choose what intention it is we want to have done it for. I think most of us really want our intentions to have some kind of impact on our decision, an impact we do not choose, but an impact that really does affect our choice of action. Thus I think that reasons should play a sort of causal role in our actions. I will grant that this intuition limits the amount of free will we would like, but as I will explain later, that is not problematic if it can be incorporated correctly.

Furthermore, it is true that the lack of either kind of contrastive explanation does not reduce the coherence of O'Connor's theory. However, this does not make its inclusion any less important. The reason no such explanation can be given is because he has declined to make reasons play any kind of causal role in our actions. Thus, this criticism is closely tied to the above criticism. I think we need to give a proper agent-causal account that allows reasons to play some sort of causal role, which in turn will give us the contrastive explanation we desire. Only then will we have a theory that is satisfactory.

So how can we give an account of free will that claims that our actions are caused by us as free agents, as well as by our reasons? Given the current debate, Clarke has provided the most successful attempt at doing this, but it, too, runs into many problems. First and foremost, he claims that we, the agent, have control over our reasons by ordering our desires, beliefs, judgments, and so on. But this doesn't seem right; we do not have ultimate control over our beliefs and desires. The lack of said control is where our bad habits we would like to give up come from. Therefore, if, as Clarke claims, these reasons play a dual-causal role with agent causation, we would not even meet the Condition of Production[*] he is so concerned with. Since he firmly believes that any account of free will must meet this condition, his own view fails to meet even his own standards. Furthermore, giving causal influence to

---

[*] Condition of Production: The agent must be in full control of the causal factors of his or her actions, and the agent must determine which action is performed

things outside of our control would detract from the attractiveness of the purely agent causal theories like O'Connor's.

So where does this leave us? A purely agent causal theory cannot satisfy some of our basic intuitions about free will and a dual-causal theory fails to achieve the goal it set out to attain. We want to be the sole cause of our actions so that our freedom to choose is without constraint, but we also want our reasons for acting to be a part of that cause. We want to have our choices be completely undetermined, but we also want to be able to give a contrastive explanation for why we acted the way we did and not some other way. These paradoxes have led some, such as Thomas Nagel, to conclude that free will is simply an incoherent idea that we should just give up on[*]. But I think that the reason these contradictions arise is because of the way determinism and indeterminism have been thus far applied. If we change the way we think about events and causal necessity, we actually can give a satisfactory account that satisfies all these conditions.

Determinism is often thought of as the idea that every event is made causally necessary by the events prior to it. Therefore, proponents of determinism would argue, if we know absolutely every fact there is to be known about the world in one instance, and the laws of nature, we can predict everything that follows because there would be absolutely no uncertainty. Indeterminism is the competing theory that is nothing more than a negation of determinism. It does not say that nothing is made causally necessary by the events prior to it, but it does say that not *absolutely everything* is made causally necessary by the events prior to it. These two ideas have been generally thought to apply to every event across time and space, but they have not taken into account time within an event. That is to say, determinism would argue that every event is predetermined, indeterminism would argue that not every event is predetermined, but they both seem to agree that an event is either determined or it is not. In other words, either an event was made causally necessary by the events prior to it, and was always that way, or it wasn't and was never that way. I propose that this line of thinking is false. It is my belief that the causal necessity of a given event's occurrence can change over time. In what follows I will explain how this is possible and what kind of impact it will have on an agent causal view.

The proposition takes the form, event E is determined at time T if it is causally necessary at T. Now determinists would rightly argue that, according to their theory, all events are always determined. This is what allows them to say that, given omniscience about the present facts and laws of nature, they can predict any event in

---

[*] Nagel, Thomas. "Freedom." *Free Will*. Ed. Gary Watson. Oxford: Oxford University Press, 2003. Print.

the future. However, my view is not ground on determinism, but indeterminism, considering I advocate for an agent causal view. Therefore, the real point this proposition has to make is about indeterminism. Instead of just claiming that not all events are determined without a qualifying statement concerning time, I will now take indeterminism to be the claim that not all E_t are determined. Therefore it may be the case that all events are determined at one time or another, but as long as they are not all determined at all times, determinism would not be true.

Let's break this down further. There are three broad periods of time that we care about with respect to causation: the past, present, and future. Determinism, with respect to these time periods, would make the following three arguments:

D_1: Every event in the past is made causally necessary solely by events prior to it.

D_2: Every event in the present is made causally necessary solely by events prior to it.

D_3: Every event in the future is made causally necessary solely by events prior to it.

Classical indeterminism, or the version of indeterminism that did not differentiate in events with respect to time, would likewise have made three similar arguments:

I_1: Not every event in the past is made causally necessary solely by events prior to it.

I_2: Not every event in the present is made causally necessary solely by events prior to it.

I_3: Not every event in the future is made causally necessary solely by events prior to it.

Now that we are allowing events to change their causal necessity through time, indeterminism has the further advantage over determinism in that it no longer requires all three arguments to be true, while determinism does. Therefore, let us now say that indeterminism is the idea that either D_1, D_2, or D_3, is false, and that either I_1, I_2, or I_3 is true[*]. Right off the bat, it seems we can make a judgment about D_3 and I_3. Events can only be made causally necessary by the events directly prior to it. When determinists claim that, given omniscience about all the relevant facts and laws of nature at time A, they would be able to predict with absolute certainty the events at any given time B after A, they do not mean that the events of A directly cause B (unless, of course, B occurs directly after A). If this was the case, you could theoretically take out all other events between A and B and B would still happen out of necessity. No, what determinists mean is that A necessarily causes A+1, which in turn necessarily causes A+2, and so on until time B. This necessary chain of causation is what makes B's causal necessity tied to A. However, since it is the events directly prior to B (B-1) that actually make it causally necessary, it is important to note that B is not

---

[*] At least one of the deterministic propositions must be false and its counterpart true for determinism as a theory to be false. However, it may be the case that more than one, or even all three propositions, favor indeterminism

actually made causally necessary until B-1 makes it so. And since events cannot exert real causal influence until they happen, B would not be causally necessary until B-1 happens. Therefore, even if the world is deterministic, events in the future cannot actually be causally necessary because the events prior to them that would have to have taken place already to make them causally necessary have not yet occurred. This leads us to conclude that D_3 is false and I_3 is true.

I am aware that there is much disagreement about whether future statements can even hold a truth value. If they cannot, then I concede that the argument for determinism will be unaffected by what was just said. Instead, it would only need D_1 and D_2 to be true to hold up. However, I must stress to determinists that, if they accept what has just been said to be true, this does affect their theory as a whole. They cannot accept that future statements hold any truth value, lest D_3 be shown false, and they cannot make the case that events in the future are causally necessary, lest they attempt to invoke D_3. What determinists may still use their theory for is to claim that all events in the present and past are made causally necessary by prior events, and, given omniscience about all facts and the laws of nature at any given time, they may predict the events that immediately follow out of causal necessity. This is, however, a more limited version than the determinist theory that currently prevails.

As for D_1 and D_2, I cannot show that they are either false or true. After all, this is what is at the core of the debate between determinists and indeterminists. But since my agent causal view requires indeterminism to be true, I will assume for the purposes of this paper that I_2 is true and D_2 is false. Most agent causal theories assume this anyway, but they also would assume that I_1 is true and D_1 is false. This is where I must part ways with O'Connor and the other agent causal theorists. I propose that both D_1 and I_2 are true, and I_3 would be true as well if it could hold a truth value. This means that not all events are causally necessary until after they have happened. At that point they do become causally necessary. I suppose another way of putting it would be that we truly do have open choices prior to and at the time of performing the action. However, once the action has been performed, it had to have been done that way. I realize this may be confusing to some people, but let's take a look at it in respect to an agent causal theory and see how they affect each other.

Unlike either O'Connor or Clarke, I think reasons play both a causal and non-causal role. Until we actually perform the action, we are presented with various reasons to make our different possible choices. While we are completely free to make whatever choice we want, the reasons each come with their own attractiveness, pulling us toward making that decision. A rational actor would always make the choice that gives him or her the most reason to

make it. For example, consider the case of Bartholomew. Bart was down on his luck lately because he was laid off from his job. The money was running low and the bills were stacking up. Fortunately for Bart, he was selected to go on a game show. Once there, during the game, he was presented with a choice of two prizes: $10,000 or $500. He could choose either one, but just one, and without any repercussion. So the only difference between the two choices was the monetary value, and in that respect, one choice was far greater than the other. Now in this case, both options had reasons to pick them with their own attractiveness. The $500 would have been an attractive choice because it would make Bart $500 richer than he currently was. However, the $10,000 choice was far more attractive than the $500 because it would just make him so much more wealthy than the $500 would. This would give Bart far more reason to choose the $10,000 than the $500, making it the unquestionably rational choice. After all, if Bart ended up consciously and intentionally choosing the $500, wouldn't we call him irrational? Thus, in this case, it would have been Bart who was really making the choice, but he made the choice he did because it presented him with more reason to make it, *and* because he acted rationally. That latter condition is important because it shows how an event can be determined after it has happened, but not before.

An event in the present, that is, an event as it is in the process of occurring, is not determined because the agent still has the full capacity to act rationally or irrationally. After the agent has made his/her decision, then the event becomes determined. In other words, we could not say Bartholomew would have chosen the $10,000 before he did so because he still had the choice to act irrationally. However, once he did choose the $10,000, we could say that it had to happen that way, that the appeal of the greater amount of money was the reason he made his decision, and therefore it caused him to make it that way. Another way of putting it is that the reasons truly did cause his action, but not until he chose them to do so. And so once it has happened, looking at it retrospectively, we can say that it was causally necessary that he chose the ten grand.

Considering this role for reasons, and the idea that events are made causally necessary only after they happen, agent causation has the ability to solve all the paradoxes of free will. First and foremost, it gives us the truly open choices we desire in free will. This comes from I_2 and I_3 (if future statements have any truth value). However, despite this freedom of choice, we can still give a contrastive explanation after-the-fact thanks to D_1 being true. Furthermore, we get the satisfaction that a purely agent causal theory gives in that our decision is caused only by us while we are making it. But we also get the benefit of our reasons playing a role in our decision, and even a causal role once our decision has been made.

I have no doubts that my view will come under much criticism. For starters, I assume most people will ask, if our only freedom is to act irrationally, isn't that a severe limit on our free will? To answer, not really. In the cases where there is a single rational choice to be made, with the other choices presenting such inferior levels of reason as to be irrational, then no it would not be a limit. In fact, the ability to act irrationally is what would give us any freedom of choice at all. If we didn't have this capability, then we would not play any causal role in our actions. Our reasons would always decide our fate and we could only do the most rational actions. Even if we don't exercise the ability to act irrationally, it is important that we still possess it, lest we throw away our freedom. Now it should be said that this only applies when there is one clear rational choice to be made. In many cases, such a choice does not exist because there are multiple options that are either comparably good or comparably bad. In those cases, the agent does not choose to act rationally or irrationally, but simply chooses the action to perform. Once the agent has acted, we can say he or she did so for whatever reasons attracted him or her to that choice, and that those reasons then caused the action, in retrospect. So it is not that acting irrationally is the only time an agent gets to exercise choice. I just chose an example that used that function first because it best illustrated the point.

To take another example that illustrates a choice between two rational options, let's consider the case of Jasmine. Much like in the example I discussed earlier to show O'Connor's theory, Jasmine was at a food court that served both hot dogs and hamburgers. The hot dogs were cheaper, but the hamburgers were more delicious. It worked out so that those two options were roughly equally attractive to Jasmine (or at least close enough that there was no clearly better choice). Until Jasmine chooses, her choice is completely and truly open. Supposing she chose hot dogs, we could, after the decision has been made, say she chose hot dogs instead of hamburgers because they were cheaper, and the fact that they were cheaper provided a more compelling reason than the hamburgers could provide through their taste. However, we could also make the opposite statement had Jasmine chosen the hamburgers instead. In other words, the reasons that are more compelling cause the agent to choose that action, but they do not become more compelling until he or she decides that they are.

Another criticism I expect to face is Strawson's self-creation argument. In response to his argument, I will agree that the kind of person you are affects your decisions. For example, if I loved cucumbers more I might be more inclined to eat them when they're presented to me because the reasons to do so would be more attractive. I will also agree that, in light of the above statement, to be truly responsible for your decisions you must, at least in part, be responsible for the way you are. However, beyond that point, I must deny him the rest of his argument. I do not

think that being responsible for who you are in any way requires an infinite regress. To account for how we may be in part responsible for the way we are, I will take a page from Robert Kane's book and use his idea of self-forming actions (SFA's)[*]. However, unlike Kane, who thought that SFA's could only occur under extraordinary situations, I propose that we always have the potential to perform an SFA. If we want to become more athletic, we can force ourselves to get into an exercise routine. If we want to quit an addiction, we can force ourselves to stop using the addictive substance. If we want to become a kinder person, we can always start making a habit to volunteer in the community. These actions will affect you later in life and, as a result, change the attractiveness of different options. However, these SFA's are not easy to do. The more unattractive an option is to us, the harder it is for us to choose to do it. Some options can be so unattractive that they are nearly insurmountable without outside assistance, such as overcoming a drug addiction. But no one ever said that changing yourself and becoming the kind of person you want to be is easy, and at the very least this theory shows how we might be able to do so. As long as we can affect who we are through SFA's, we can be responsible for the kind of person we are and the decisions that result.

Finally, complete skeptics such as Pereboom might still deny this theory because science seems to indicate that agent causation simply doesn't happen. But, given the specific agent causal account I have advanced, there is no real way to determine its truth or falsity scientifically. After an event has happened, it will work out so that it does not look like it could have happened any other way, therefore giving no indication that agent causation was involved. However, the truth of the matter is that it would have looked just as certain even if the outcome was different. Even modern neuroscience and its ability to map certain brain functions cannot disprove agent causation because it is not an event itself. Since it would affect other processes without being affected itself, it is impossible to conclude that it doesn't play some influence in determining brain states. Thus, I find no reason to grant Pereboom the claim that agent causation does not seem to be accurate simply because science does not seem to point towards it.

But, so far, these objections have all been common to agent causal theories in general. To defend my theory in specific, I shall now answer some objections raised by our fellow philosopher, Mason Westfall. The first

---

[*] Kane, Robert. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Free Will*. Ed. Gary Watson. Oxford: Oxford University Press, 2003. Print.
Kane, in this paper, attempted to provide an account for free will in an indeterministic world without agent causation. His idea was that, while most events were actually determined, during special and extreme occasions of decision making, we are pulled in two different directions between the choices. At this point, we have control and can use free will to make the choice. Since these occasions are extreme, they tend to be momentous decisions that will shape our personality and thus alter our future behavior. For Kane, it was through these SFA's that we gained some semblance of responsibility for and control over our actions. Though I am not using his strict idea of SFA's in my paper, I owe credit to him for originating the idea.

objection he raised was toward my point that future events, if they can hold any truth value, must not be causally necessary. He argued that, for events A, B, and C, where A makes B causally necessary and B makes C causally necessary, C would indeed be a necessary consequence of A. I will concede that, they way he has phrased the point, he is absolutely correct. However, this does not accurately represent the way I have phrased my argument. This causal line has semantically indicated that it all takes place in the present tense. If it were to reflect future causation, it should be considered as three events, A, B, and C, where A makes B causally necessary and B *will* make C causally necessary. But it is important to note that this indicates C is not yet causally necessary because B cannot make it so until it happens. Of course, he has also brought up the idea that time block theorists would reject this entire notion because all events are equally real in their view. I again concede this point. However, if I am not mistaken, time block theory also presupposes determinism. Obviously, agent causal theorists do not accept determinism as true, so our theories already preclude the possibility of time block theory. This is a conflict inherent to not just my theory, and not even just agent causal theories, but all theories of free will that rely on indeterminism. So it should not be seen as a problem for my view in specific. Furthermore, I already conceded that determinists who reject what I argued about future events, for whatever reason, would not have their theory affected by it. My goal is not to show determinists that they are wrong, for I do not think this can be done. I just wish to show that, assuming they are wrong, an acceptable version free will is possible.

A second objection he raised concerns the point I made that we can have free will through the choice to act rationally or irrationally. The issue in contention is whether rationality is actually a choice. But to say that I argued rationality is always a choice is to misunderstand my case. There are many instances in which acting rationally or irrationally is not under our control. I personally, can attest to this through my arachnophobia. I know that when I see a spider, the level of harm it can do to me is vastly disproportionate to the fear it strikes in me and the reaction I have to it. Yet I act in this irrational manner anyway. However, this does not mean that we can never consciously make the choice to act irrationally. In the sense that acting rationally is acting in accordance with reason, and that irrationality is disregarding reason and acting in a manner that will clearly not realize the most benefits for us, then yes this can be done willingly on occasion. The example I gave earlier of Bartholomew I believe clearly illustrates this point.

A third objection Mason raised was that, given the theory I proposed, reasons do not actually cause action. He supported this objection by saying that, since reasons for a given choice would still exist even if the choice they

would have caused does not, they cannot be the things that cause the choice when it is actually made. This objection, I think, stems from a failure to accurately comprehend my position. Given the standard view where events were either always causally necessary or else they never were, this would be true. However, under my view, events do change from being indetermined to determined. So while it is true that the reasons would exist regardless, they change the nature of their existence depending on the agent's choice. Given that, under this view, events become causally necessary after they happen from the reasons that caused them, by very definition of the position the reasons do cause the action. Therefore, given the way I have set the view up, I do not think the existence of reasons as providing causes can be questioned. It would be fair, however, to simply question how much to believe this view as a whole for reasons external to it. Obviously, if you have trouble seeing reasons in this manner, this view will probably not present an acceptable possibility for you.

Finally, Mason has suggested that, despite the fact that neuroscience cannot actually prove agent causation to be wrong, it can offer enough alternative explanation to render agent causal theories obsolete. Since this is combined with the lack of agent causal theories' ability to empirically prove themselves correct, science still renders a problem to those theories. Allow me to start by saying I wholly agree with this. As science progresses it may very well gain the capacity to empirically explain all aspects of free will we are currently debating on the philosophical level. It may even turn out that agent causation is false. However, until it does so, we must consider every plausible possibility. Furthermore, if agent causal theories are correct (as those of us who propagate them tend to think), then it will happen that science will never be able to explain neuroscience to an acceptable degree that provides the kind of free will only agent causation can offer, since the agent causal aspect will be unobservable. As to this second part of the objection, that the theory really isn't provable, this is an acceptable limitation. Again, it is not the goal of this paper to prove any and every other theory wrong, but just to offer a new kind of view that can actually show free will to be possible in the manner we feel it should be.

As a closing thought, some might object to this view on the grounds that I am attempting to "have my cake and eat it too" by using the advantages that both determinism and indeterminism offer, while avoiding both their drawbacks. However, as I laid out through my discussion of the paradoxes of free will, any account that can properly satisfy all our desires for free will must necessarily run into this issue. So the real question I would ask to those who would raise this objection is, "Is there any account of free will you would accept? Is this a substantial reason to discredit a potential solution?" I leave answering that question up to you.

It seems as though an agent causal view that allows events to become determined after they have happened is able to resolve the paradoxes of free will, and resolve most of the criticisms other such views run into. Considering all the advantages this view has, it would seem logical to seriously consider changing the way we look at determinism. After all, the idea that an event must either always be determined or never be determined seems to be the very thing that has caused most of the problems we have with free will today.

MAKING SENSE OF GLAUCON'S CHALLENGE

— Brian Joo —

Introduction

The eudaimonist worry about justice is how an other-regarding virtue can somehow contribute to an agent's own state of happiness or well-being. If we want to know whether the presence of justice in a person's life will make them happier than without it, we need to first know what justice is. This is the challenge raised by Glaucon (and Adeimantus). Socrates is requested to state what justice is and how it is desirable for its own sake and not merely for its consequences. However, justice by itself, without referring to its consequences, is supposed to benefit the agent by making them happier than being without it. But benefits are a kind of consequence, and those are specifically requested not to be included in praising justice in itself. Is Glaucon's (and Adeimantus's) challenge incoherent? I argue that it is not if consequences are properly interpreted and understood. Unfortunately, there is no higher standard to which we can appeal to for knowing which interpretation of consequences is correct. I present three different interpretations by: 1) Christopher Kirwan and Terence Irwin, 2) Robert Heinaman and Nicholas White, and 3) Andrew Payne. I assess each of their merits in terms of how well they account for Glaucon's particular request and how coherent and plausible it is for Socrates to in fact praise justice for its own sake without having to refer to *consequences*.

Glaucon's (and Adeimantus's) Challenge
Tri-partition / Taxonomy of Goods

Glaucon begins by dividing up goods into three classes so as to determine which class Socrates thinks justice belongs to (357b-358a)[*]. They are: 1) Intrinsic goods valued for their own sake, such as joy. 2) Goods valued for their own sake and for their consequences, such as health and knowledge. 3) Instrumental goods valued not for their own sake, but for their consequences, such as money and medicine. Socrates declares that justice belongs to the second class, "the finest of goods" (358a). Glaucon wonders how this is the case when the general consensus is that justice is itself onerous, only to be practiced for the sake of the rewards and popularity that comes from a reputation for justice (358a). So he requests Socrates to prove how justice can be valuable in and of itself. Moreover, he wants Socrates to do so without depending upon its beneficial consequences. I will explicate Glaucon's request in order to illustrate the peculiar demand placed on Socrates so we can later assess its coherency.

---

[*] G. M. A. Grube, *Plato's Republic* (Indianapolis/Cambridge: Hacket Publishing Co., 1992). All future references to the Republic will be from this source in Stephanus pagination, unless stated otherwise.

Why People Practice Justice

Glaucon plays the role of the devil's advocate and assumes the position of the general public's conception of why one should act just, in order to demonstrate that justice is in fact not an intrinsic good, but strictly an instrumental one. He assumes that suffering from injustice is far worse than benefiting from it (358e-359a). Since most people cannot act unjustly with impunity, they come to the rational agreement to be just towards each other in order to secure the benefits of other people being just towards them without anyone having to suffer from injustice (359a-b). I reason to be just because the justice of others is conditional on my own.[*] For example, if I always borrow money from people with the intent of never reimbursing my lenders even when I told them I would, for a time I will reap the rewards of others while not sacrificing any goods of my own. However, I will eventually develop an undesired reputation of not being trustable, making it incredibly difficult to maintain a reasonable relationship with anyone else for any given purpose. Therefore, the life where I end up suffering from injustice is far worse than the benefits I reaped from doing so. It is in one's self-interest to practice justice as an instrumental means to avoid being mistreated by the injustice of others.[†] This, Glaucon argues, is the nature of justice and its natural origins (359b).

The Ring of Gyges

Having established why people practice justice, Glaucon wants to argue that anyone that is allowed to do what they truly desire with impunity will not be just. Rather, they will naturally practice injustice because it is the nature of one's interest to "outdo others and get more" (359c). Those who practice justice do so unwillingly, as something necessary because they lack the power to do injustice with impunity, not as something good for the sake of itself.[‡]

Glaucon tells the story of the Ring of Gyges to illustrate the counterfactual instability of justice being good for its consequences (359d-360d). Gyges the shepherd comes across a magical ring imbued with the power of invisibility which he can command at will. Wielding such power, Gyges immediately turns to a life of injustice, satisfying all his desires of wealth and power: he seduces the king's wife, kills the king with her aid, and establishes himself as the new ruler of the kingdom. The ring has not only granted Gyges the power of invisibility, but also the power of undetectability. That is, Gyges acquired the power to be unjust with impunity. Furthermore, it also allowed

---

[*] David Brink, "Handout #8: Justice and Politics in Plato's Republic", p. 3. (FA'10 – Phil 161)
[†] This is the rational egoism argument.
[‡] This already starts to highlight what kind of good a class-III good is. Justice is onerous, pursued for its instrumental benefits of alleviating people's shortcomings of being unable to practice injustice with impunity. We can draw an analogy to the pursuit of medical procedures, which are also considered onerous. They are not pursued for the sake of themselves, but for their instrumental benefit of alleviating some pain that is suffered by the individual that cannot be alleviated through their own will.

him to act unjustly without the reputation of injustice. Gyges can remain reaping the benefits of injustice while avoiding all of its negative consequences. The story shows that anyone who *can* practice undetected injustice *will* naturally do so rather than practicing justice. It seems then that unfettered injustice results with more desirable consequences. Therefore, it seems that practicing injustice with impunity is what truly contributes to one's eudaimonia.

Appearing Just vs. Being Just

If undetected injustice is what is actually preferred by all people, then the *appearance* of justice seems to be what is valued. Extending the example used earlier, if I were to continually borrow money from lenders, never reimburse them, and somehow maintain a reputation of accountability and responsibility that was admirable to the point of envy, then lenders would have no reason not to believe that I will reimburse them and will continue to do so even though I have the strict intention to do otherwise. It seems then, that it is in the person's best interest to act unjustly while *appearing* to be just so as to reap the maximal amount of benefits from others while sacrificing little to nothing of one's own behalf. This shows that the conventional defense of justice as a rational agreement of self-interest mentioned before is unstable. That is, one does not need to practice justice to appear just. Therefore, there is no reason to act justly, since acting unjustly with both impunity and a reputation for justice results in a better situation.

"Two Lives" Scenario

Glaucon's final provision expands on the previous idea that what people actually desire is to *appear* just, rather than to *be* just. He introduces a scenario where one is offered the choice between two lives (360e-362c). On the one hand, you can choose the life of perfect injustice, where one always avoids the negative consequences of acting unjustly coupled with the greatest reputations and awards normally offered only to the very just. That includes wealth, societal influence, respectful authority, popularity, and other valuable consequences. On the other hand, you can choose the life of perfect justice, coupled with an unearned and unfair reputation for injustice, along with being deprived of all the rewards that accommodate being just. Glaucon suggests that when offered such choices, no one would prefer the life of perfect justice coupled with the reputation for perfect injustice since that life is doomed to unfairly suffer "every kind of evil" (361e). Socrates is challenged to show how the worst just world is always better than the best unjust world. If Socrates can succeed in providing a counterfactually stable defense of

justice, he will demonstrate that justice is good in itself, independent of its entailed consequences.[*] By doing so, he can show that justice is in fact a class-II good, valued for both its own sake and its already conceded consequences.

<div align="center">Coherency of Glaucon's Request</div>

It is unclear how to understand this other-regarding virtue when Socrates is supposed to praise justice in a counterfactually stable way without referring to its consequences. Is Glaucon's (and Adeimantus's) demand coherent? Can justice be shown to contribute to the agent's eudaimonia without having to resort to any of its benefits? Can justice be genuinely praised in and of itself? As a class-II good involving both intrinsic and instrumental benefits, his request on the face of it seems incoherent. But we *can* make sense of this if we are to specify the beneficial consequences into different kinds. For example, we can understand how justice is to benefit the agent possessing them while distinguishing those benefits from the ones more traditionally tied to its practice, such as reputation and reward. The former kinds of benefits is identified by David Brink as *intrinsic* while the latter as *extrinsic*.[†] Intrinsic benefits can be specified as those benefits in which by virtue of possessing them, it contributes to the best kind of life we can lead, while extrinsic benefits are the benefits which are acquired by living the best kind of life. Understood in this way, the intrinsic benefits constitute the extrinsic benefits. Health is an instructive example for illustrating this distinction. Health is praiseworthy in its own right, where we do things for the sake of health (e.g. exercise, diet, etc.). At the same time, health enables us to live the best kind of life we can lead (e.g. sex, play, dance, etc.).[‡] So both health and justice, under the intrinsic-extrinsic distinction described above, above, has no problem in satisfying Glaucon's taxonomy of a class-II good. More to the point, his challenge *can* be rendered coherent.

<div align="center">Different Interpretations of Coherency</div>

But now we run into another sort of problem: how are we exactly supposed to interpret the beneficial consequences so as to correctly render Glaucon's challenge coherent? This, I think, is the genuine problem in

---

[*] It is worthwhile mentioning how Adeimantus adds to this provision. He mentions how people consider unjust deeds as being more profitable than just ones (364a) and how such wealth and power can allow for greater persuasion of the gods' judgment so as to avoid their punishments for being unjust (364b-366d). In other words, being unjust results with wealth and power, but also upsets the gods. However, I can use my wealth and power to praise the gods and sacrifice to them so as to please and persuade them from being upset by my unjust acts. So I can be unjust and escape even their punishment. So Adeimantus also requests of Socrates to explain how justice benefits its possessors and how injustice harms them without referring to its consequences. He adds at the very end, "Don't, then, give us only a theoretical argument that justice is stronger than injustice, but show what effect each has because of itself on the person who has it – the one for good and the other for bad – whether it remains hidden from gods and human beings or not" (367e).
[†] Brink, "Handout #8", p.3
[‡] In no particular order.

understanding the coherency of the challenge made against Socrates. How we understand what beneficial consequences are determine how we are to make sense of Glaucon's challenge, which in turn influence how Socrates's defense of justice will look. In other words, if the challenge is altered, then so is the response. Extrapolating what Socrates's defense of justice does (or can) look like is beyond the scope of this paper. Instead I will be limiting my focus strictly on the different interpretations of Glaucon's challenge. In particular, I will focus on the two more famous interpretations that have been offered, namely Kirwan & Irwin's take and Heinaman & White's take. Moreover, I will use Andrew Payne's recent interpretation to address the flaws he points out in both interpretations, along with providing his solution that avoids such flaws.

Kirwan and Irwin Interpretation

One interpretation, offered by Kirwan and Irwin, involves the rejection of relying on any of the beneficial consequences for Socrates's defense of justice. Rather, the only benefits of justice that are allowed to count are those constituted by the state of being just.[*] For example, justice is understood as an essential component of happiness where justice is necessary but not sufficient for happiness. How being just directly benefits our happiness as a constituent of our happiness, Kirwan and Irwin thinks, is the proper way of understanding Glaucon's request.[†] Reputations of justice and rewards for being just are not to count because they are not constituted by justice, but are the results of what is constituted by justice. They are valuable consequences, but related to justice in an indirect way. This interpretation, Irwin argues, allows for a clear distinction between class-II and class-III goods by ruling out the valuable consequences of justice and instead focusing on the constitutive, intrinsic benefits.[‡]

One major problem with this restrictive interpretation, Payne thinks, is how Glaucon incorporates "valuable consequences" into his model when illustrating the kind of praise of justice he desires. In particular, the scenario involving the two lives (of the unjust best world vs. the just worst world) includes one life, namely the unjust life, having all sorts of valuable consequences, including wealth, popularity, social connections, etc. Payne argues that if Glaucon's praising of injustice includes valuable consequences as Kirwan and Irwin understand it, then it is implied by the model Glaucon provides that Socrates may include at least some of the consequences of justice when praising

---

[*] Christopher Kirwan, "Glaucon's Challenge," *Phronesis* 10 (1965), pp.162-73; Terence Irwin, *Plato's Ethics* (New York: Oxford University Press, 1995), pp.181-93

[†] There is an essential similarity to the aforementioned intrinsic-extrinsic distinction offered by Brink. The intrinsic beneficial consequences of justice are supposed to constitute and allow for the best kind of life worth living, while the extrinsic benefits are those consequences that result from that best kind of life. So for Kirwan and Irwin, what will count are the intrinsic benefits of justice.

[‡] Irwin, *Plato's Ethics,* pp.190-1

justice for its own sake.[*] Otherwise, the challenge would be stacked against Socrates in a way that would make it impossible for him to respond. If someone asked me a math question using numbers, but requested me to compute it without numbers and furthermore respond without numbers, the challenge would appear just as stacked and incoherent. Kirwan and Irwin mistakenly exclude valuable consequences when a proper solution would include it.

Heinaman and White Interpretation

Heinaman and White, on the other hand, prefer a less restrictive interpretation that allows for a subset of the beneficial consequences of justice.[†] The subset that White allows for involves those consequences which are directly caused by justice's own power that is independent of other factors, as opposed to consequences produced involving other distinct factors. Heinaman allows for more leniency in what is to be included in the subset. That is, he allows for Socrates to include in his praise of justice all beneficial consequences except those dependent on a reputation for being just. They both rely on the causal relation between being just and its directly resulting benefits as their main criterion for inclusion into the subset.

A major problem with Heinaman and White's account, as Payne points out, is that this special mode of causal consequence is nowhere to be found in the *Republic*.[‡] If we are to instead understand the causal relation in a more intuitive, primitive way, then there would be no way to adjudicate between class-II and class-III goods under this standard. For example, physical exercise would be causally related to its product health analogously to justice being causally related to one of its beneficial consequences. Then physical exercise would also be praised as something good in itself, which intuitively is not the case. A similar problem threatens Heinaman's interpretation. If we are to consider something as a class-II good because of its valuable consequences apart from any based on its reputation, then a class-III good such as medical treatment would also have to be considered a class-II good since it produces valuable consequences independent of any resulting from its reputation. The main problem with Heinaman and White's interpretation is how it does not preserve the distinction between class-II and class-III goods. Without clarification of this distinction, we would not know where to place justice. So Glaucon's request of Socrates to praise justice strictly as a class-II good would be impossible to satisfy. What we want is to interpret consequences in such a way where our intuitions match up with how we categorize goods under Glaucon's taxonomy.

Payne's Interpretation

---

[*] Andrew Payne, "The Division of Goods and Praising Justice for Itself in *Republic* II," *Phronesis* 56 (2011), pp.74-5
[†] Robert Heinaman, "Plato's Division of Goods in the *Republic*," *Phronesis* 48 (2003), 309-35; Nicholas White, "The Classification of Goods in Plato's *Republic*," *Journal for the History of Philosophy* 22 (1984), pp.393-421
[‡] Payne, "The Division of Goods…," pp.75-7

Payne provides us with a new interpretation that allows for the inclusion of valuable consequences in the praise of justice while preserving the distinction between class-II and class-III goods. He argues that the special subset of valuable consequences that Socrates is allowed to appeal to when praising justice as a class-II good is to strictly involve the sort of consequences which "testifies to the good working order of the soul that is just."[*] He calls this subset of consequences *criterial* benefits of justice, which is to be distinguished from what he calls *fringe* benefits of justice.[†] The former is supposed to provide a kind of standard (or criterion if you will) for evaluating when that thing (in our case justice), which the criterial benefits are of, is a good instance of its type. The latter is understood as those benefits that do not provide such a standard, and so, for example, cannot indicate proper functioning of a just soul. Only the former, criterial benefits of justice are to be accounted for when praising justice, never the latter. The difference in kinds of benefits also preserves the distinction between class-II (criterial and fringe) and class-III (fringe) goods.

I think that Payne's use of criterial and fringe benefits provides an interesting and promising way to interpret Glaucon's challenge, including the particular conditions he has provided. Let us take for instance the two-lives scenario. So Glaucon distinguishes between the worst just life and the best unjust life. He requests Socrates to praise justice in such a way that the former is always preferred over the latter. The latter includes valuable consequences in its description, while the former is stripped of them. Moreover, the former, being stripped of its rewards and reputations, is requested to be praised without relying on what has been stripped. This is a critical problem in the coherency of Glaucon's request since it is difficult to understand how to praise justice without including some of the valuable consequences which has been amply included in praising its opposition, namely injustice. Payne's analysis of criterial and fringe benefits can easily explain this so as to eliminate the incoherency.

We can distinguish between the activities of an unjust man and a just man along with the criterial benefits of those activities which can indicate to us the success of the relevant activity. For the unjust man, maintaining a reputation of justice is a criterial benefit because such a reputation is required in order for the unjust man to avoid the penalties of his deeds. On the other hand, a reputation for justice is not an indicator of the success of the just man. The just man's primary activity is being just, not simply appearing just sufficient for retaining the reputation for it. Therefore, the reputation for justice is not a criterial benefit for the just man, but instead is a fringe benefit.

---

[*] Payne, "The Division of Goods…," p.60
[†] Payne, "The Division of Goods…," p.60

The same can be said of the benefits of wealth, power, respect, and so forth.[*] So when Glaucon does appear to be guilty of incoherency when requesting Socrates to praise justice without relying on those same valuable consequences included in Glaucon's previous praising of injustice, he is innocent. The same valuable consequences can be understood as different kinds of benefits, with kinds being relative to whether one is considering a just or an unjust person. [†]

Conclusion

Can we make sense of Glaucon's (and Adeimantus's) challenge of praising justice in itself without referring to its consequences? I definitely think we can, so long as we define consequences in such a way where we can make a distinction between benefits that are directly and indirectly related to justice, where the former constitutes the latter. However, how we are supposed to define "directly" and "indirectly" is less clear (and very fascinating). Here we can have numerous interpretations, the merits of all of which are only to be assessed by comparison with other interpretations in terms of how closely it captures the core idea of praising justice in itself while adhering as closely as possible to the *Republic* II. I have assessed three separate interpretations, and of them I am in most favor of Payne's account of criterial and fringe benefits. It avoids the problems of the other two, which were excluding all use of valuable consequences in praising justice and failing to preserve the distinction between class-II and class-III goods. Until problems in Payne's account have been identified and another, less problematic account can be substituted for its place, I think Payne's account is the best (in the Popperian sense where best is meant as the least falsifiable).

---

[*] These same benefits would also be considered fringe benefits for the unjust man. Since appearing just is itself sufficient for a reputation for justice (which is required for a successful life of injustice), then all other benefits of wealth, power, respect and so on would be consequential benefits of the successful life of injustice.
[†] Payne, "The Division of Goods…," pp.71-3

FEYERBEND'S ARGUMENT FOR THEORY PLURALISM REVISITED
Hidden Facts & Alternative Theories

― Max Mergenthaler ―

Thoughts without content are empty, intuitions without concepts are blind.
Imanuel Kant CPR, A 51

Abstract

Paul Feyerabend is known for representing the claim that theoretical pluralism is to be preferred over theoretical monism in science. For, he claims, that there are potentially refuting facts that become only available with the invention of alternative theories. I will argue that Feyerabend's argument is wrong. (I) I will begin by briefly exposing Feyerabend's argument for his claim that alternative theories increase the empirical content of existing theories and should therefore be desired. Then I will argue that this argument has serious flaws. My criticism of Feyerabend will consist of four parts. First (II), I will show that Feyerabend conception of scientific refutation bears some problems, which can be solved, however, with an alternative formulation that will be presented. Secondly (III), I will evaluate his view of the Brownian Particle as potential refuter of classical thermodynamics, and show (IV) that Feyerabend is wrong in thinking that it supports his theory. Thirdly (V), I will attempt to show that if we distinguish the refuting character of facts and the actual refutation, Feyerabend theory might hold well. Fourthly (VI), I will evaluate Couvalis' attempt to rescue Feyerabend's thesis from Laymon's attack on the hidden fact thesis and conclude that it fails Laymon's point since both views are compatible. Finally (VII), I will present Worral's logical attack on Laymon's and Couvalis's conclusion that states that a successful alternative was needed to refute thermodynamics.

(I) In most of his work, but specifically in his *Problems of Empiricism* (1965), Feyerabend claims that there are some anomalous or potentially refuting facts that will remain necessarily concealed if there are not alternatives to the theory to be tested. We will refer to these facts that remain *blind* as hidden. According to him *"theoretical monism"* is to be rejected because it leads to scientific stagnation and non-empirical dogmatism. This amounts to saying that the empirical content of a theory depends on its rivals (cf. Worral 1978).

In his main argument for the convenience of theoretical proliferation in science, one of the major premises, besides his account of the invariance of meaning, states the following: In the absence of theoretical pluralism or proliferation some potentially refuting facts will remain forever hidden, reducing drastically the amount of possible criticism and leading to unwanted dogmatism (cf. Ronald Laymon 1977).

Insofar, Feyerabend can be read as claiming that for every theory T, the knowledge of an alternative to T is a necessary condition for the discoverability of some experimental anomalies for T. As Feyerabend writes, this is equivalent to claiming that not all experimental anomalies for T can be discovered in the absence of another theory

T', which is an alternative to T. [*] This means that certain facts remain necessarily hidden in the absence of alternative theories that unearth them.

Although his paper seems to lack a solid argument to support the thesis that significant facts cannot be discovered in the absence of an alternative theory, Feyerabend appeals to the case of the scientific difficulty that the justification of correct and relevant descriptions of Brownian motion[†] represents.[‡] This description could, according according to him, not be determined from the available observation data that was at hand, so that... "[...] the actual refutation was brought about [...] via the kinetic theory and Einstein's utilization of it in the calculation of the statistical properties of the Brownian motion" (Feyerabend 1965:176).

Without going into more detail on how exactly this happened, Feyerabend goes on to generalize the given example in the following way: "Consider . . . a theory T that successfully describes the situation in the domain D. From this we can infer (1) that T agrees with a finite number of observations (let their class be F); and (2) that it agrees with these observations inside a margin [E] of error only. Any alternative that contradicts T outside F and inside [E] is supported by exactly the same observations and is therefore acceptable if T is acceptable (we shall assume that F are the only observations available)" (Feyerabend 1965:173)"

That F exhausts the available observations should suggest that we ought to state the critical question: Should scientists try to invent alternative theories given that T agrees with all available observations "within a margin [E] of error"? Feyerabend's answer is "yes", since we otherwise fall in an unwanted dogmatism that ignores

---

[*]"Not only is the description of every fact dependent on some theory (which may, of course, be very different from the theory to be tested), but there also exist facts that cannot be unearthed except with the help of alternatives to the theory to be tested and that become unavailable as soon as such alternatives are excluded. The answer to the . . . question is, simply, no. Consider what the discovery of the inconsistency between the Brownian particle and the phenomenological second law would have required. It would have required (a) measurement of the exact motion of the particle in order to ascertain the changes in kinetic energy plus the energy spent on overcoming the resistance of the fluid, and (b) precise measurements of temperature and heat transfer in the surrounding medium in order to ascertain that any loss occurring here was indeed compensated by the increase of the energy of the moving particle and the work done against the fluid. Such measurements are beyond experimental possibilities. Neither is it possible to make precise measurements of the heat transfer; nor can the path of the particle be investigated with the desired precision. Hence a "direct" refutation of the second law that considers only the phenomenological theory and the "fact" of Brownian motion is impossible. (Feyerabend 1965:175-176)"

[†]As most introductory books to physics explain, Brownian motion is the presumably random movement of particles suspended in a fluid a liquid or a gas. It was named after Robert Brown, who noticed in 1827 a "rapid oscillatory motion" of microscopic particles within pollen grains suspended in water. The effect, being independent of all external factors, is now ascribed to the thermal motion of the molecules of the fluid. However the first satisfactory theoretical treatment of Brownian motion was made by Albert Einstein in 1905. Jean Perrin experimental studies provided verification for Einstein's mathematical formulation and are regarded as one of the most direct verifications of the kinetic-molecular theory of gases. Before Einstein Brownian motion represented an anomaly for classical thermodynamics.

[‡] "As is well known, the Brownian particle seen from a microscopic point of view is a perpetual motion machine of the second kind, and its existence refutes the phenomenological [thermodynamics] second law. It therefore belongs to the domain of relevant facts for this law. Now could this relation between the law and the Brownian particle have been discovered in a direct manner, i.e., by an investigation of the observational consequences of the phenomenological theory, without borrowing from an alternative account of heat? (Feyerabend 1965:175)"

relevant refuting facts.

According to Feyerabend, the case of the motion of Brownian particle satisfies the above condition of compatibility with available data and can be generalized in the following way to show the necessity of an alternative theory. We will refer to this generalization as the *generalized refutation schema:* "[…] assume that a theory T has a consequence C and that the actual state of affairs in the world is correctly described by C', where C and C' are experimentally indistinguishable. Assume furthermore that C', but not C, triggers or causes, a macroscopic process M that can be observed very easily and is perhaps well known. In this case there exist observations, viz., the observations of M, which are *sufficient* for refuting T, although there is *no possibility whatever* to find this out on the basis of T and of observations alone. What is *needed* in order to discover the limitation of T implied by the existence of M is another theory, T' which implies C', connects C' with M, can be independently confirmed, and promises to be a satisfactory substitute for T where this theory can still be said to be correct. Such a theory will have to be *inconsistent* with T, and it will have to be introduced not because T has been found to be in need of revision, but *in order to discover* whether T is in need of revision. (Feyerabend 1965: 176) (my italics)"

(II) We will start our critique of Feyerabend by pointing out that this version of the generalization is apparently incoherent. For how can there possibly exist observations M that refute T but are not possible to find out on the basis of T and observation alone? If the observation of M is sufficient to refute T, then T must entail the negation of M. In his 1978 paper John Worral (1978:303) wields a devastating analysis[*] of the logical implications of this version of the *generalized refutation* that shows that "despite its air of depth" this description of how theories get refuted is either incoherent, empty or a bizarre version of the Duhem-Quine thesis.

However if one follows the principle of charity, the *generalized refutation schema* can be modified and developed to a "non trivial interpretation". If one takes, as Laymon does, that "observations" means "observation sentences", then T being by T' given M is correctly described by claiming that T' entails some observation sentences M_O that T does not. T is then refuted just in the sense that the success of T' in explaining M via M_O is a reason for rejecting T, since T' explains *more.* Here refuting facts are hidden in the sense that one cannot know if an observation that is not

---

*For a full length refutation see Worral (1978:303). Here the condensed version. 1.) According to Feyerabend a theory T entails C where C is false but not testable. 1.1) C' is correct. 1.2) C and C' are inconsistent. **So** T is not refutable by C' because C and C' are not distinguishable through experiments. **However** 2.) T' changes things because T' entails C' ; since C' causes an observable M. This means that T' $\rightarrow$ (C' $\rightarrow$ M). [Since T' $\rightarrow$ C' we follow that T' entails M too]. 3.) M is true, so M confirms T'. **But** 5.) T is not affected by M unless 5.1) T entails 'C $\rightarrow$not- M'. If that's the case then T can be refuted without need of T'. **or** 5.2) T' entails C $\rightarrow$ not-M. [This makes T refutable by appealing to T']. **So** 5.2 allows us to derive new testable consequences M for T from T'. **however** 6.) T and T' are inconsistent: T $\rightarrow$ C and T' $\rightarrow$ C' . **THUS** T' entails 'T $\rightarrow$ P' for any statement P.

explainable by T is explainable by an alternative that does not exist (cf. Ronald Laymon 1977).

Put briefly, Laymon's non trivial interpretation of the *generalized refutation schema* goes something like this:

a.) T entails that if certain initial conditions *I* are satisfied then M_O will be observed. b.) Whether or not the conditions *I* really hold are not being settled by scientists.

c.) Nevertheless M_O is regarded as experimentally false. **Thus** d.) Finding out if *I* holds would increase T's empirical content and T would be in fact refuted.

d.) T' predicts that if *I* holds, then M_O* will be observed. e.)If we assume that M_O* is observed then this gives us inductive reason for believing that *I* holds. Finally, f.) if *I* is indeed accepted, then since 'T & I' entails the experimentally false M_O, T would be refuted with a modus tollens (cf. Worral (1978:304).

This modified *generalized refutation schema* will concern us later. Back to Feyerabend, Laymon makes us aware to the fact that Feyerabend substitutes the talk of "refuting facts" for the talk of just "facts" or even "the relevance" or "refuting character" of facts in the next passage:

Both the *relevance* and the *refuting character* of many decisive facts can be established only with the help of other theories that, although factually adequate, are not in agreement with the view to be tested. (Feyerabend 1965:176, my italics)

Those no longer hidden facts can not be anything but the observations of M, which get their relevance by the alternative Theory which is able to explain them. Equally what gives them a refuting character is that T cannot explain them. Laymon concludes: "Therefore, it is no longer the facts themselves that are hidden in the absence of an alternative but merely their character" (Laymon 1977:229). This will show to be decisive when analyzing Couvalis' (1988) attempt to rescue Feyerabend.

The concrete instantiation of the *generalized refutation schema* for the Brownian motion case could look as follows: T stands for phenomenological thermodynamics and T' for statistical mechanics. M stands for the [movement of] Brownian particle (cf. Feyerabend 1965:176). Due to the different nature of both theories with C and C' things get more complicated, for our goals it suffices to say that C and C' are what we called initial conditions *I* (Laymon 1977: 230).

(III) Ronald Laymon argues that an examination of this case shows nevertheless that no alternative theory

"is or historically was thought to be necessary in order to justify descriptions of Brownian motion that "directly" refute thermodynamics", rendering Feyerabend appraisal incorrect (Laymon 1977). To prove this let us consider the following three questions:

1) "Did the refutation of thermodynamics require the presence of the alternative statistical mechanics?

2) Did statistical mechanics render visibility to the otherwise invisible fact that Brownian motion is a perpetual motion machine of the second type?

3) Does the case of Brownian motion instantiate the generalized refutations schema in any of the interpretations considered?" (Laymon 1977:233)

Let us begin our critique with Feyerabend's claim that, since the needed measurement of motions, temperature and heat transfer are "beyond experimental" possibilities, a "direct" test of the phenomenological second law is not possible. This seems wrong, as Laymon argues, if we for example take Kelvin's version of the law[*]; where the apparently impossible measurements "of the exact motion of the particle" are not required because "any work performed as a unique final result is sufficient to refute the second law" (Laymon 1977: 234). This means that if the medium of the Brownian particle is isolated and of uniform temperature, any single particular motion would refute the law.

So, the only way in which Laymon can make sense of Feyerabend is by interpreting him as saying that the required measurements are beyond experimental possibilities because its impossible to know if the particle is properly isolated and if its motion is indeed the only final result using the usual measuring instruments. However he dismisses this over-restrictive interpretation immediately because "direct" means for Feyerabend without consideration of another theory of heat. So, understood correctly, Feyerabend's claim is simply incorrect, since "there was direct [without employment to an alternative theory] use made of Brownian motion as a *test* of thermodynamics" (my italics) as the history of science proves.

Laymon claims that the observations of the movement of the Brownian particle, once Auxiliary Hypothesis that made the movement compatible with Thermodynamics appealing to Mill's method of concomitant variation†

---

[*] A transformation whose only final result is to transform into work, heat extracted from a source which is at the same temperature throughout is impossible. Qtd. in (Laymon 1977: 234)

† Symbolically, we could represent the method of concomitant variation as follows: (± represents a shift):
A B C occur together with x y z
A± B C results in x± y z.

were ruled out, show that Thermodynamics did not require an alternative theory to unearth problematic facts. This is not equivalent to the claim that the rejection of thermodynamics was empowered by an alternative theory, we will see later that this confusion led Couvalis (1988) to claim that Laymon's account was mistaken. We will now return to the historical argument.

There were many other candidates for explaining the unusual motion of the Brownian particle in a way consistent with the second law of thermodynamics in terms of local disturbances.[*] However, these candidates were tested and excluded via the method of concomitant variation ending with the french physicist Gouy using these failures of explanation to 'reject' the second law. According to Laymon (1977:236) Poincare supported Gouy's in his rejection. However, it is important to note that Poincare claims to find just the second and not the first law problematic, while Feyerabend seems to be claiming that making precise and refuting measurements out of the range of experimental possibilities is necessary if both laws are to be tested.

(IV) We will not go into the details of the history of science and will assume this historical account is correct. But, does it fit the modified *generalized refutation schema*? For Laymon the answer is no. The formalized modified schema goes like this. We begin analyzing compatibility of a theory with observation.

1. "$I(x) \& T' \rightarrow \varphi(x)$"   and   2. "$I(x) \& T \rightarrow \text{not-}\varphi(x)$"

Where 'I' are the initial conditions (in this case 'I(x)'= 'x in a closed system in equilibrium') and '$\varphi(x)$' = ' is some proposition about the mean displacement of x'.

Laymon notes that historically both 'I(x)' and its negation were compatible as descriptions with the available observations concerning temperature, electromagnetic, and mechanical disturbances making the choice between 1. and 2. or T and T' very complicated (Laymon 1977:231). However, Feyerabend's schema is supposed to offer a solution. But lets note first that 'not-I(x)' will not yield $\varphi(x)$ given T or T'. Moreover, 'not-I(x)' "is not sufficiently quantitative speaking to yield on the assumption of T a description $\psi(x)$ of the motions of the Brownian Particle that is as specific and quantitative as $\varphi(x)$" (Ebd:232). Lets also consider that a Description D(x) that was compatible with 'not-I(x)' *and* observation such that

3. $D(x) \& T \rightarrow \psi(x)$ where $\psi(x)$ is compatible with observation and is as specific and quantitative as $\varphi(x)$

———————————————————————
Therefore A and x are causally connected
[*]   Laymon 1977 mentions unequal temperatures in strongly illuminated water, evaporation, air currents, heat flow and capillarity.

was historically *not* available. This means that an ordered pair {D(x), ψ(x)} compatible with observation did not exist at the time. Here *compatibility* is related to the then held observation sentences about the Brownian particle (cf. Ronald Laymon 1977:232).

Lets now turn to refutation, i.e. the success of T' over T. The difference here is that an ordered pair {D(x), ψ(x)} namely {I(x) φ(x) exists such that 'D(x) & T'→ ψ(x)'. The explanatory failure of T and the success of T' is a good inductive reason for thinking that:

a.) No acceptable {D(x), ψ(x) } exists for T

b.) I(x) and φ(x) are true (those are only compatible with available observation. So have inductive reason to believe that I(x) holds)

Thus, with Modus Tollens on 2. the assumption of I(x) leads to 'not-T', thus refuting T appealing to T'. Summarizing: refutation occurs when we consider, grounded on the success of the alternative theory T', both 'I(x)' and ' φ(x)' as true and the refuted theory has the form "I(x) & T → not-φ(x)".

We can see that the *generalized refutation schema* does not fit the presented historical facts. For, the essential sub-conclusion of the schema was that:

{D(x),ψ(x)} does not exist such that D(x) and ψ(x) are compatible with observation and such that 'D(x) & T → ψ(x).'

Where this non-existence, as Laymon (1977:239) explains, must be grounded in the success of the alternative theory T' and the failure of T to find an acceptable {D(x),ψ(x)}. But as we showed before, scientists argued without the necessity of using alternative theories for the nonexistence of the mentioned ordered pair {D(x),ψ(x)}. Mill's concomitant variation, was used, to show that the state of affairs describable by {D(x),ψ(x)} did not exist for T, thus making T *at least* problematic. This shows that *problematic facts* were available without an alternative theory. However this is neither equivalent to the claim that problematic facts are sufficient to refute a theory, nor to the claim that scientists rejected the theory in light of some anomalies.

Given that all causes that could explain the particle motion consistently with Thermodynamics were unsatisfactory, we can criticize Feyerabend thesis of hidden facts, because neither the knowledge nor the acceptance of statistical mechanics (or any other alternative theory) "is or was thought to be necessary in order to (inductively) see Brownian motion as a counter instance to the second law of thermodynamics" (Laymon 1977:240).

(V)   We saw that the success of the alternative theory of statistical mechanics was, contrary to what Feyerabend claimed, not necessary for finding refuting facts of thermodynamics. But we could still ask, did or could the success of statistical mechanics have helped the demise of thermodynamics? i.e.: "does the predictive success of statistical mechanics reinforce the refutation of thermodynamics in a way that instantiates the pattern of Feyerabend's generalized refutation schema?" (Laymon 1977:240).

And here the answer is 'yes'. Lets analyze the situation in light of the modified general refutation schema. First, the condition that a substantial range of descriptions of the initial conditions that is compatible with observation and is available is satisfied. With thermodynamics and the available observations back then it was impossible to decide whether the Brownian particle was in a fluid that was in thermal equilibrium or not. Let's remember that this is not unimportant because it means that T could not have been refuted by a modus tollens just recurring to available descriptions. The second condition was also satisfied, as there were no other descriptions of the initial conditions compatible with observation known at the time, such that Thermodynamics could yield a successful and precise quantitative prediction of the movement of the Brownian Particle. For we only had the hypothesis that states that if the initial conditions of the experiment are taken to be that the fluid environment of the Brownian particle is suffering local disturbances, then the Brownian particle will suffer motion. "This consequence does not, however, give a quantitative description to this motion". So here, statistical mechanics "enjoyed a clear advantage" since the "simple description of thermodynamic equilibrium yielded a specific and successful quantitative prediction" (Laymon 1977:241).

(VI)   In his 1988 paper George Couvalis offers what he thinks is a severe criticism of Laymon's unjust attack on Feyerabend, in which he claims to show:

 "that the Second Law of Thermodynamics was not refuted before the Kinetic Theory's predictions were confirmed, and that it could not have been refuted without the confirmation of the remarkable predictions of some rival theory." (Couvalis 1988: 415)."

In Couvalis' view Feyerabend argues that it is impossible to discover directly that the Brownian movement of particles violates the Second Law of Thermodynamics and that it was only when the predictions made by critics of the Kinetic Theory were tested that the Second Law was refuted. Although this seems to be exactly contrary to what we have been arguing against, it will turn out that both Laymon and Couvalis argue in a very similar manner, although in a different tone. What Laymon claims is not that Thermodynamic was completely refuted due to the

anomalies that could not be properly explained, in fact he denies the possibility of this theory being refuted by a simple modus tollens without the exact and precise predictions that statistical mechanics made available. Laymon is arguing against the radical epistemic claim that inconsistent facts, which have a refuting character, are in principle unavailable in absence of an alternative theory.

Laymon claims a.) that the facts are not really unavailable and Couvalis claims that b.) an alternative theory that gives, precise and exact predictions of those anomalies, and accounts for more phenomena in a better way is to be preferred over its rival. Logically both claims are compatible, and historically it seem that this was also the case.[*]

Strictly speaking Laymon is *not* arguing that a careful examination of the Brownian case "shows that it neither is nor was historically thought to be necessary to confirm Einstein's predictions of the behavior of Brownian particles in order to refute the Second Law of Thermodynamics" (Couvalis 1988:415). I will now discuss briefly Couvalis' arguments and show that they do not really contradict Laymon's interpretation. Couvalis' first argument is that the experiments using the method of concomitant variation that were carried in the early 20th century did not refute the second law but just made it suspicious.

The first point I want to make is that the section of Poincare's article that Laymon says supports Gouy's claims does nothing of the kind… Poincare is saying that if the Brownian motion does not borrow anything from an external source of energy then Carnot's principle (the Second Law) is violated. He is being noncommittal as to whether the Second Law is violated." (Couvalis 1988:416).

In this case it is obvious that Laymon's point against Feyerabend's hidden facts is accepted by Couvalis. And both agree that due to lack of a quantitatively better theory thermodynamics was not completely rejected by the scientific community. Nevertheless, Poincare does not argue that the second law is violated, he just claims that further experiments should be carried out to determine if all possible causes for explaining Brownian motion were excluded, rendering the law refuted. Clearly there were relevant facts available even in the absence of an alternative†.

---

[*]   For example, consider the case of Mercury in the late 19th century. Astronomers observed that Mercury's orbit did not fit Newton's predictions. However the deviation of Mercury's orbit wasn't taken to be a definite refutation of Newton's theory until Einstein's theory provided an alternative explanation for the deviation. Here the potentially refuting fact was unearthed long before the alternative theory was even conceived. Couvalis historical account of the brownian particle, which supposedly contradicts Laymon's will turned out to be analogous to the example offered.

†   This again does not mean that the theory is refuted, given the Duhem-Quine problem of auxiliary hypothesis that have to be ruled out. For, as it is well known, and we have explain above, there were other possible explanations for the problematic facts and this makes it logical impossible to completely refute a conditional that has a disjunction as an antecedent with a modus tollens.

The second argument of Couvalis is somewhat stronger and claims that these experiments *could* only have provided grounds for suspicion and not for rejection, which again would be compatible with Laymon's account. He argues: "The method of concomitant variation could not have been used to refute the Second Law on its own because the source of the Brownian motion could well have turned out to be due to the action of an unknown energy source. This is why most eminent physicists waited until experimental evidence confirmed the remarkable predictions of the Kinetic Theory before they were willing to commit themselves to the view that the Second Law was wrong." (Couvalis 1988:418)

Again, here it is clear that Couvalis is arguing that only either clear predictive failure or the confirmation of the remarkable predictions of a rival theory is sufficient to refute a well-confirmed theory (cf. Ibidem). However we arrived at the same conclusion with Laymon when we explained that the possibility of refuting T with a modus tollens was not given until T' was available rendering I(x) plausible. Thus both seem to agree that the Second Law could only be completely rejected by the confirmation of the predictions of some rival theory such as Einstein's version of the Kinetic Theory. However, this is not unproblematic, as we will now see.

(VI)   Let us finish with one last consideration regarding the modified *generalized refutation schema* and the general possibilities of refutation in light of the Quine-Duhem thesis. We saw that an alternative theory can facilitate the rejection of another theory. This however rests on allowing the statement I & 'not-M_O' as a possible falsifier of T. However as Worral (1978:308) notes, *I* is a very high level statement, in that it implies that one takes into account all possible sources of thermodynamics disequilibrium that could cause the Brownian particle to move. Worral asks us to substitute Laymon's *I* (some system is closed and in equilibrium) with "something more like a genuine statement of initial conditions" *I'* that says that particular forces and particular sources of disequilibrium have been avoided and add a particular assumption *A* that states that these potential disturbances are the only ones in operation. However as Duhem pointed out those statements are never testable in isolation. So scientists have to choose the best available auxiliary hypothesis *A*. If we consider that in our case *A* was tested by "winning out against various proposed competitors", and that therefore Thermodynamics was a Theory that was "capable of being consistent with experimental results only when conjoined with other conjectures about possible perturbations, initial conditions, […] etc. *and* when conjoined with the best auxiliary conjectures then available this theory *was inconsistent with experimental results.*" (Worral 1978:307, my italics). With Worral we have to conclude that thermodynamics was as logically refuted as any scientific theory can be before the invention of statistical

mechanics.

We have shown that Feyerabend's thesis that facts can often not be unearthed without an alternative theory thus lowering the empirical content of predominant theories can not be seen as a good argument for Theory pluralism. Furthermore we should distinguish between the question of the availability of relevant facts that pose a refuting character and the question of the actual dismissal of a theory in light of an alternative. Laymon shows that Feyerabend is wrong because refuting facts are indeed available in absence of an alternative, while Couvalis showed that a better and alternative theory seems to be fundamental for the rejection by the scientific community of a theory that cannot account for serious anomalies. However, if we just consider the logical possibilities of refuting a theory then Laymon and Couvalis are mistaken in believing that statistical mechanics unearthed hitherto hidden motivations for refuting thermodynamics. However, if we consider the history of science as presented by Laymon and Couvalis, Worral is mistaken in ignoring the fact that until it's statistical-kinetic rival offered quantitative successful predictions about the Brownian motion, Thermodynamics was not considered refuted and therefore rejected. We are confronted with the following question: while doing philosophy of science, should be prefer a logical or historical account that takes scientific practice into consideration? Which answer should be regarded as correct has to remain an open question in this paper.

Works Cited

Problems of Empiricism, Paul Feyerabend, in R. G. Colodny (ed.), Beyond the Edge of Certainty. Englewood Cliffs, N.J.:Prentice-Hall,1965.

Feyerabend, Brownian Motion, and the Hiddenness of Refuting Facts, Ronald Laymon,*Philosophy of Science,*Vol. 44, No. 2 (Jun., 1977), pp. 225-247

Feyerabend and Laymon on Brownian Motion, Spyridon George Couvalis,*Philosophy of Science,*Vol. 55, No. 3 (Sep., 1988), pp. 415-421

The Theory-Ladenness of Observation,Carl R. Kordig, *The Review of Metaphysics*Vol. 24, No. 3 (Mar., 1971), pp. 448-484

Review: Feyerabend on Observation and Empirical Content, Elie Zahar,Reviewed work(s): *Der Wissenschaftstheoretische Realismus und die Autorität der Wissenschaften* by Paul Feyerabend,*The British Journal for the Philosophy of Science,*Vol. 33, No. 4 (Dec., 1982), pp. 397-409

Is the Empirical Content of a Theory Dependent On Its Rivals?*, John Worrall : in I. Niiniluoto and R. Toumela (eds.): 'The Logic and Epistemology of Scientific Change', Acta Philosophica Fennica, 30, 2-4, pp. 298-310*

THE RATIONAL AREN'T RICH

— Brad Segal —

In this paper I will present a variation of Newcomb's problem to bring out how, faced with a Newcomb problem, a causal decision theorist will undertake an action with the best possible payout (given their circumstances) while an evidential decision theorist will be presented with better options and will make—arguably—a less rational decision. Through this I will show how the Newcomb problem is one in which a rational agent will simply not be presented with the options that can lead to the best possible outcome. Thus I will argue that the prominent protestation, "if you're so smart, why aren't you rich?" should be set aside as an objection to causal decision theory.

Here is a variation of Newcomb's problem. Imagine that two people, Randy and Ira, are walking down a street on a Monday morning. They see their mutual world-class psychiatrist standing, arms outstretched, with $1,000 of cash in each hand. He stops them, presents each man with $1,000 and asks, "Do you want to take this money? I made a prediction for each of you last Friday. If I predicted that you would turn down this $1,000 then I deposited $1,000,000 into your bank account. If I predicted that you would take the $1,000 then I put nothing into your bank account." He presses each man to decide. "Take or leave the $1,000?"

	Predicted to take cash	Predicted to refuse cash
Take Cash	$1,000	$1,001,000
Refuse Cash	$0	$1,000,000

Table 1: Possible payout matrix. Actions are in the column to the left and the possible states of the world are represented in the row above the possible payouts.

Both men have heard that the psychiatrist occasionally offers this exact ultimatum to his patients and are aware that he has been 95% correct in his predictions. Though initially a little surprised, Randy thinks for a second and then takes the $1,000. Ira decides not to take the money. The psychiatrist tells them that he had predicted correctly in both cases. There was $1,000,000 transferred into Ira's bank account on Friday while Randy had received nothing. As they walk away Randy asks Ira, "Were you not thinking? The money was already in your bank account, why didn't you take that extra $1,000?" Ira responds, "If you're so smart, why aren't you rich?"

This situation illuminates the crucial difference between evidential and causal decision theorists. Evidential decision theory recommends *auspicious* acts correlated with favorable results. James Joyce sums up the point of view by stating that their actions are "evaluated in terms of the *evidence* they provide for thinking that desirable outcomes will ensue" [1]. Conversely, causal decision theorists base their actions on *efficacy*, acts that have a causal role in securing a benefit. Joyce argues that rational agents should seek to maximize efficacy [1]. Two events that are causally connected will be correlated, but correlation does not entail causation [2]. Thus causal decision theorists seek to perform acts that will cause a positive state, not that will just be correlated with a positive state. A decision for them should bring about the desired outcome, not just be a sign of it. Given a decision, usually evidential and causal decision theorists will agree. But in cases like Newcomb's problem, when there is a correlation between an agent's act and the outcome though no causal connection, the two theories can recommend different actions.[*]

Based on the psychiatrist's prediction and his decision, Ira appears to be an evidential decision theorist. If he were to take the $1,000 it would be correlated with the psychiatrist's prediction that he would take the cash and thus a sign that nothing had been deposited into his bank account. But Ira prefers having a million dollars put into his bank account to having nothing deposited. His refusal of the cash and the prediction that he would refuse the cash are correlated. In turning down the cash, he is seeking the good news that the money was indeed deposited. Refusing the cash, he reasons, would be auspicious evidence that the psychiatrist had put a million dollars into his bank account.

But before examining Ira's notorious comeback, we should scrutinize Randy's first question, why did Ira not take the extra $1,000? The psychiatrist's prediction was made the previous Friday. At the time of Ira's decision, he had no *causal influence* on the psychiatrist's prediction. David Lewis points out how the prediction could be made before, simultaneously with or even after the decision to take or reject the $1,000 [3]. The point is that Ira's decision has no causal influence on whether or not there is $1,000,000 in his bank account. Though the decision and act of taking the $1,000 are correlated with the outcome of having a million dollars in his bank, the money would be there (or not) regardless of what he does. From Randy's perspective, at the time of their decision on Monday, the probability (if one were to take the $1,000 there would be $1M in his bank) = probability (if one were to refuse the

---

[*] It should be noted that Randy and Ira need not know that they are either causal or evidential decision theorists; they could just be following their guiding intuitions telling them what to do. In this scenario the two decision theories would describe each of their decision-making processes.

$1,000 there would be $1M in his bank) because his decision cannot change Fridays prediction. Likewise the probability (if one were to take the $1,000 there would be $0 in the bank) = probability (if one were to refuse the $1,000 there would be $0 in the bank). So Randy raises a legitimate question because the state of Ira's bank account is independent of what Ira decides to do. In the situation that Ira was in, he made a decision between refusing the money and ending with $1,000,000 or taking the money and having $1,001,000 because in either case the state of the world was such that the money was already deposited into his bank. So, as Randy subtly conveyed, Ira irrationally turned down $1,000.

Yet Ira retorts with, "If you're so smart, why aren't *you* rich?" Awfully bitter for someone who just made a million bucks, no? Well Randy was not presented with the same options that Ira was. The psychiatrist did not predict that Randy is the type of person who would turn down the money and thus there was no money deposited into his bank account the previous Friday. Randy knows enough decision theory to understand that, at the time of his decision, he would have no causal influence on whether or not the money had been deposited. Sure, he might have been able to influence the psychologist's prediction before Friday. But even if the money had been deposited, Randy knows that he should still take the $1,000 because he understands that on Monday it would be irrational to turn down the cash as that he has no causal influence on the state of his bank account. It is precisely because the psychologist knew that Randy understood this that the money was not deposited. Possibly the psychologist knew that *Ra*ndy was *ra*tional and would follow the recommendation from causal decision theory. It would be even more compelling to take the $1,000 if Randy was aware that the psychologist knew Randy to be rational and a follower of causal decision theory.[*] So it is not Randy's fault that he is not rich, but given the circumstances he found himself in, he actually did the best that he possibly could have because he avoided walking away with $0. And if he was presented with Ira's options, he would have ended with $1,000 more than Ira had.

After Randy calmly explains all of this, Ira smugly presses back, "Don't you wish you were like me then?" True, Randy would be rich if he had been presented with Ira's options. But even if he had been like *Ira* (who he thinks of as quite *irra*tional) and was presented with Ira's options, it still would have been better to take the $1,000. And precisely because Randy understands that taking the $1,000 is always better due to the act's causal

---

[*] It could be that Randy has never heard of causal decision theory. Randy needs to only have some understanding of his decision making and need not be able to spell out the details of causal decision theory for Randy to predict the psychologist's impression of himself.

independence from Friday's prediction and its efficacy in bringing about the best possible result given the circumstances, he never would have been presented with Ira's options in the first place.[*]

If Joyce is correct in asserting that a rational agent should perform acts that causally bring about the best result, taking the extra $1,000 would be the most rational decision. So the fact that Randy isn't rich is not the fault of causal decision theory, it is a manifestation of his rationality. As Lewis points out, Newcomb's is a problem where irrationality is richly rewarded [4]. There is no requirement that a rational agent will always end with more money than an irrational agent. Instead it is quite easy to imagine that, precisely *because an agent is rational*, there are certain situations in which a rational agent will not be presented with the options that lead to the best possible outcome. This just appears to be one of those cases. But this is not a reason to abandon causal decision theory. Instead, it seems like a reason for causal decision theorists to avoid Newcomb problems (though in this case they will walk away with $1,000 which, considering their circumstances, is the best possible outcome).

Yet Randy may have the last laugh yet. Once an evidential decision theorist really *understands* the argument for causal decision theory, it seems like they should be almost contaminated by it. Let me explain; an obvious way to be presented with Ira's options is to live as an evidential decision theorist. However, imagine that a better-studied Ira was offered the $1,000 on the sidewalk a month after his first opportunity. If he really understands the argument behind taking the cash, he should realize that his decision on that Monday morning really does have causal independence from the prediction made on the previous Friday. He may then realize that it would be pointless to refuse the $1,000 in front of him because he now has no role in the psychiatrist's prediction and, thanks to his loud and jubilant attestation of his status as an evidential decision theorist, the money is probably already in the bank account. But the realization that he is inclined to take the money would serve as evidence that the money had not been deposited into his bank account. And furthermore Ira would have—at some point after understanding causal decision theory—realized that he had doomed himself to come to his apprehension about his decision's causal independence at the moment when he has to decide. Thus he had destined himself to be inclined to take the $1,000, further evidence that the money was not deposited. Even if he decides to play it safe and refuse the money, once he studied the theory, how could he trust that the psychiatrist didn't expect him to flip at the time of his decision? Then he would end with nothing! No, it would seem much safer to take the $1,000.

---

[*] Unless (1) Randy lived a convincing lie in front of the psychiatrist before the preceding Friday—which is implausible because it would defeat the purpose of going to a psychiatrist or (2) if the psychiatrist had been flat-out wrong in Randy's prediction.

Works Cited

Joyce, James. *The Foundations of Causal Decision Theory*. 1st. Cambridge: Cambridge University Press, 1999. 1-268. Print.

Weirich, Paul, "Causal Decision Theory", *The Stanford Encyclopedia of Philosophy (Fall 2010 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2010/entries/decision-causal/>.

Lewis, David. "Prisoners' Dilemma is a Newcomb Problem." *Philosophy and Public Affairs*. 8.3 (1979): 235-240. Print.

Lewis, David. "'Why Ain'cha Rich?'" *Noûs*. 15.3 (1981): 377-380. Print.

MODERATE-LIBERTARIANISM[*]
A Moderate Dose

― Mason Westfall ―

In this paper I will consider Michael Otsuka's attempt to reconcile a libertarian right to self-ownership with equality. This reconciliation is often called left-libertarianism. In order to do this I will first sketch Robert Nozick's argument against redistributive taxation on the basis of a libertarian right to self-ownership. I will then introduce Otsuka's distinction between a right to self-ownership and a stronger conjunctive right to self-ownership and ownership of worldly resources. This will lead us to a discussion of the disagreement between Nozick and Otsuka on the correct theory of justice in acquisition, which is the substantive disagreement between them. I will then consider gifts and bequests, which pose a serious challenge to Otsuka's theory. This will lead us to a consideration of how Otsuka's argument schema can be co-opted to defend a wide variety of moral views. Finally I will propose a middle way, attempting to avoid the problems posed on one side by Nozick's argument, and on the other by gifts and bequests. Ultimately I will argue that my proposal should at least be considered a very strong competitor, and that the more general lesson to take from Otsuka is that the libertarian right to self-ownership is a remarkably pliable right, which can be fit in to a plethora of distinct moral theories, and as such tells us very little by itself as to what constitutes a just state.

Nozick argues that redistributive taxation is morally indistinguishable from forced labor on self-ownership grounds (Nozick 169). Since I have a property right in myself, I am entitled to decide what I will do insofar as this does not infringe on such a right (or other basic rights) of others (this is what the right to self-ownership amounts to for Nozick). But part of this right to self-ownership is a right to the fruits of my labor (Otsuka 60). If, after I have provided for my own bare necessities, I choose to spend my surplus time in leisure, watching a sunset at the beach or some such, it would certainly be wrong for the state to force me to work extra hours to provide for others who cannot provide for themselves. If, however, I choose to spend my leisure time (time over and above the time I use to provide myself necessities) working extra hours such that I may have access to various luxury goods, many think it is permissible to tax this surplus income for redistributive ends (Nozick 170). But in both cases I am exercising my

self-ownership rights in accordance with my preferences. To seize my income equivalent to five extra hours of work is just as much a violation of my self-ownership rights as is forcing me to work an additional five hours during my leisure time. The only difference between the two cases is in one I prefer surplus income and in the other I prefer surplus free time. But surely this isn't a difference that makes a difference. Why should contingent psychological facts about my preferences affect your respecting my self-ownership rights? So it seems either both forced labor and redistributive taxation are permissible, or neither are. Given Nozick's strong commitment to self-ownership, it is obvious to him that neither is permissible. They both constitute violations of individual rights.

Otsuka argues that this is only correct so far as it goes. In a small subset of cases he simply grants Nozick his conclusion, but then argues that the conclusion only extends to the majority of cases with the help of a suppressed and dubious premise that he is happy to deny. In order to make this point, Otsuka points out that in the majority of cases, self-ownership is insufficient to acquire goods. Instead, one must mix one's work (exercise one's self-ownership) with worldly assets (land is a paradigm example) and derive surplus profits from this mixture. But these cases are notably different. While one's right to self-ownership might be considered straightforward, a property right in worldly goods is notably less straightforward. But in order for Nozick's argument against redistributive taxation to go through, he needs property rights in worldly goods. In fact, Otsuka thinks he needs assent to the general premise that "one's right of ownership over worldly resources that one uses in order to earn income is as full as one's right of ownership over oneself." (Otsuka 75) But this premise seems intuitively false. Seizure of my land, though it may be a serious violation of my property rights, is a much less severe violation of my rights than violation of my body. Nozick might respond that he does not endorse that property rights are *as* stringent as the right to self-ownership, but that they are both genuine rights, and as such cannot be violated.* Exactly how this gets cashed out, the salient point for Otsuka is that Nozick's argument against redistributive taxation relies not on a right to self-ownership alone, but requires a further claim about ownership of worldly objects. This leaves us room to grant Nozick his right to self-ownership without his conclusion against redistributive taxation, if we don't buy the suppressed premise about property rights.

But surely we can enjoy property rights in worldly objects, at least some of the time. And when we enjoy such rights, we are entitled to the profits thereof. Otsuka grants this, but offers a relatively strict egalitarian theory

---

* It is actually not so clear that such a response is open to Nozick. He endorses a very strong notion of rights as side-constraints. This means that rights are totally inviolable. As such, it is hard to see in what sense a right could be more or less stringent, since by the very nature of the term, it is maximally stringent. This might naturally be seen to lead to an objection concerning conflicting rights, but it is beyond the scope of this paper to consider this in detail.

of property rights. This is his vehicle to reconciling a robust right to self-ownership with egalitarianism. In order to do this, Otsuka argues that Nozick's theory of justice in acquisition is inadequate, and an egalitarian theory is more plausible. Both Nozick and Otsuka are interested in Locke's theory of justice in acquisition. According to Locke, one can come to own an unowned worldly resource by "mixing his labor with it." (Nozick 174) Supposing this to be true (Nozick raises some very interesting questions about it that are beyond the scope of this paper), it is clearly insufficient as a theory of just acquisition. This is because one person's appropriating of some quantity of resource changes the position of all others with regard to that resource. As stated, it might be just for me to gain ownership of all the water on a desert island by mixing my labor with it in the right sort of way, and then charge the other inhabitants of the island a lifetime of servitude for minimal access to my water, becoming an unchallengeable despot in accordance with justice. But this is certainly not a just acquisition. This motivates Locke's proviso that there be "enough and as good left in common for others." (Nozick 175) Both Nozick and Otsuka think that this proviso is right (this is independent of any commitment to Locke's broader theory of acquisition), but a lot hangs on how it is to be understood. Nozick's proviso reads:

You may acquire previously unowned land (and its fruits) if and only if you make nobody else worse off than she would have been if she had instead been free to gather and consume food and water from this land and make use of it but not to privatize the land itself. (Otsuka 78)

Otsuka thinks this formulation is too weak, since there are cases in which one could act in accordance with it, yet disadvantage others (Ibid). Generally stated, the problem is such a formulation is entirely insensitive to opportunity costs (Otsuka 79). For this reason, Otsuka favors what he calls an "egalitarian proviso," namely:

You may acquire previously unowned worldly resources if and only if you leave enough so that everyone else can acquire an equally good share of unowned worldly resources. (Otsuka 79)

Otsuka understands "equally good share[s]" as shares that allow equal opportunities for welfare (Otsuka 80). This reading of the egalitarian proviso has the consequence that those who (through no fault of their own) are less efficient converters of resources to welfare will be entitled to greater shares of resources. This theory of property rights will demand considerable redistributive action by the state, while still respecting a robust right to self-ownership. There are two reasons for this. First, it will be in practice quite difficult to tell ahead of time what native differences in efficiency in resource conversion exist. This demands ex post facto tinkering, taxing those who make themselves better off than the mean for the benefit of those who fail to achieve the mean.   Second,

intergenerational resource transfer will not guarantee preservation of equal opportunity for welfare. In this way, it will be permissible for the state to be highly redistributive without doing anything akin to forced labor.

Unfortunately, such a strongly egalitarian conception of property rights has some rather implausible results. Specifically, if we accept the egalitarian proviso, we are committed to very restrictive requirements on gifts and bequests. According to Otsuka, these sorts of activities would be permissible "only under the extreme condition that nobody derives any net benefit from them." (Otsuka 91) On his picture, I could freely bequest my property upon my children, so long as they get a proportionally smaller share of initial resources, with similar taxation of gifts. But this seems to simply rule out bequests and gifts, since there can hardly be an incentive for my giving you a gift aside from my wish to confer a net benefit upon you. Even wholly sentimental gifts (gifts with no market value) will presumably be taxed, since knowledge on your part of my care for you will constitute a net benefit. This is highly problematic (and Otsuka is aware of it). However, the problem as I see it has nothing to do with Otsuka's project as such, but instead with egalitarianism generally. Any moral theory that demands equality will be committed to "leveling down," of which the gift problem is a species. The general worry is in a world in which I'm twice as well off as you (presumably a world in which I am not a philosophy major), the egalitarian will be committed to saying that it would be morally better to make me half as well off without changing your welfare at all. This will achieve the egalitarian goal of equality, but only at the cost of a good deal of lost and uncompensated for welfare. So the fact that Otsuka's proposed system is susceptible to such a worry seems to me evidence that it is a truly egalitarian system, instantiating both the costs and benefits thereof.

This might be soluble by considering one's propensity for receiving gifts as part of one's capability for welfare. Such that people who are particularly likable, and therefore likely to receive a greater proportion of gifts over time, will thereby have less initial resources. However, such a proposal is not without its own costs. If your equality of welfare depends on my lavishing you with gifts, there appears a burden on me to do so. I become obligated to give you gifts, which is, some might say, antithetical to gift giving.

We might think this worry can be solved by adopting a starting gate egalitarianism, such that initial conditions must conform to equal opportunity for welfare, but divergence is allowed thereafter. Society allows people to squander their opportunities. While this will solve the intragenerational worry, the intergenerational worry remains. While it is now permissible for me to give you a gift without it being negatively compensated for, it is still not permissible for me to leave my property to my child uncompensated. This is because while you and I, as

members of the same generation,* presumably had equal initial positions, and so any divergence as a result of my giving you a gift is permitted, the bequest works differently. While the money I make is not a part of my initial conditions, it will be a part of my child's initial conditions. But those conditions must be equal, so any bequest must be compensated against. In fact it seems to me that the problem extends beyond mere bequests. Such a theory will limit, not only bequests, but also any expenditure I make on behalf of my child while I'm still alive. If I wish to purchase music lessons for my child, this too must be compensated against, since many people are in initial positions that are not so pleasant as to include complimentary music lessons.† So I think these sorts of leveling down concerns are endemic to even the more moderate forms of egalitarianism, and as such should be seen as a cost to be weighed against the benefits of the theory.

Perhaps something has gone awry. Upon encountering Nozick's argument against redistributive taxation, we saw that it relied on a particular conception of justice in property rights. We then substituted Otsuka's egalitarian theory of property rights and saw that we indeed were no longer saddled with Nozick's argument, but now were unable to give a satisfactory account of gifts or bequests. Perhaps the problem is upon jettisoning Nozick's theory, we docked at the first harbor that presented itself, namely Otsuka's egalitarian proviso. But these are surely not the only theories with any plausibility. If we adopted some third theory, we might avoid both the problems that face Nozick, and those faced by Otsuka.

In fact, Otsuka has given us a schema for reconciling the libertarian right to self-ownership with any end-state distributional scheme. All one has to do is give a theory of property rights such that acquisition of property is just only if it is conducive to the favored end-state. This can then be conjoined with a robust self-ownership right and yield the favored end-state (or, more realistically, yield something distinct from the favored end-state in which redistributional state activity in order to achieve the favored end-state is permissible, and likely required). It is rather short work to construct a utilitarian proviso that only allows acquisition of property if it maximizes utility, a Rawlsian proviso that only allows acquisition of property if it is to the greatest advantage of the least advantaged, a prioritarian proviso that only allows acquisition of property if it maximizes welfare, with extra weight given to the welfare of the least advantaged, and so on. It should be stressed that adopting one of these provisos, while achieving

---

* This is, of course a highly simplified notion of generations, but it is heuristically useful.
† One might object that the starting gate is in fact earlier than childhood, say infancy. This seems to me to make starting gate egalitarianism a rather vacuous theory, since in actual fact infants are more or less equal (controlling for negligent parents). The real inequalities only start showing up once kids are going to schools of differential quality, and given access to various extracurriculars inaccessible to others. Any serious starting gate egalitarian should include these as part of the starting gate.

the ends prescribed by the relevant view, will not in fact be the end-state views in question. To take the utilitarian proviso as an example, to conjoin it with the right to self-ownership, as Nozick understands it, is not to be a utilitarian. This is because utilitarians do not recognize inviolable rights like we are imagining the right to self-ownership is (this is true even for so called utilitarians of rights). Instead, someone assenting to this view will have a broadly speaking deontological view, it just so happens that the particular rights we have bring it about that utility is maximized. Those with superficially conflicting intuitions that people have robust libertarian rights, but also that some particular end-state is just, will find the related reconciliation very attractive.

We might also consider a version of the Lockean proviso not tethered to any particular end-state. Perhaps (learning from Otsuka's objections) we might have a proviso that allows acquisitions only if the benefits outweigh the lost opportunities. This would rule out a savvy businessman appropriating all worldly resources and compensating others so they are marginally better off, while still allowing gifts. However any such principle will need much more precise spelling out, among other things, it must answer whether bequests are permissible or not. If yes, then the principle will allow rampant inequalities over time, and if no, it will place very stringent and draconian restrictions on how we may treat our children.*

The proposed view will attempt to slip between the horns of the above dilemma. The answer will be that bequests are permissible but in a limited manner. More precisely, bequests and expenditures on one's child would be permissible only insofar as the benefits conferred on the child do not outweigh the competitive disadvantages others will suffer as a result. This would result in relative freedom for vacations, and other sorts of expenditures that notably increase quality of life while providing almost no competitive advantage.† While placing very stringent restrictions on differential education, which provides huge competitive advantages. This might be a fruitful middle road between leveling down objections and rampant inequalities. I have of course not shown this view to be uncontroversially correct. There are many other objections that might be raised against such a view, and how to weigh competitive advantage against welfare increases will need to spelled out in some more detail. It will also have to compete against intuitive end-state reconciliations. However I believe I have motivated the view as at least a very strong competitor among the possible reconciliations.

---

* I am becoming convinced that bequests are an extremely problematic moral conundrum. I have the paradoxical intuitions that a parent has a right to use her fairly gotten resources to benefit her children, but her children cannot be benefited in this way. This runs into deep issues concerning the permissibility of partiality in morality that should be discussed in great depth somewhere else.

† One might argue that "being more worldly" actually does provide some competitive advantage. This strikes me as true in a small number of situations, and I think on balance the competitive advantage is quite small relative to the gain in utility.

So where have these considerations left us? If one had a pet end-state theory before entering this debate, reconciling it with a robust libertarian right to self-ownership might be an attractive option (if you are willing to pay the cost in parsimony of adding rights to your ethical theory). If, instead, one is uncommitted, I think I have shown the view that balances benefit against competitive disadvantage of others to be a strong competitor, avoiding some of the unpleasant consequences of both Nozick and Otsuka's views. However these debates shake out, Otsuka's considerations have shown that Nozick's right to self-ownership is entirely insufficient to deny redistributive taxation. And given the breadth of potential theories that are consistent with it, it actually tells us very little about what constitutes a just state. This is quite an important accomplishment in its own right, and should be treated as such.

Works Cited

Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic, 1974.

Michael Otsuka. *Philosophy & Public Affairs,* Vol. 27, No. 1 (Winter, 1998), pp. 65-92