Boris Kment, "Free Will and Ultimate Explanation." Final version published in *Philosophical Issues, 27, Metaphysics, 2017* (doi: 10.1111/phis.12103).

FREE WILL AND ULTIMATE EXPLANATION

Boris Kment
Princeton University

When reflecting on the causal antecedents of human action in the right manner and context, it is easy to get the impression that morally relevant freedom is impossible. I call this the "non-existence impression." The objective of this paper is not to determine whether this impression is correct, but to understand which features of an action's causal history generate it. According to the most familiar answers, the impression results from the observation that every human action is either (i) caused, or fully determined, by factors beyond the agent's control or (ii) the outcome of a chance process, combined with the assumption that that is inconsistent with freedom. I argue that this cannot be the full explanation, since there are possible cases of human action in which neither (i) nor (ii) is true, but which nevertheless give rise to the non-existence impression. On my alternative account, the impression rests on the implicit assumption that freedom requires involvement in the ultimate explanation of one's action (in a novel sense that I explain), and on the observation that agents do not meet this condition. It follows from plausible assumptions that satisfying this condition is impossible, since no action can have an ultimate explanation. This in turn is not a distinctive feature of human agency, but an instance of the general truth that no fact is ultimately explicable.

Let "free will" signify the kind of freedom of choice and control over one's actions that is required if one is to be morally responsible for what one does. Even though free will is arguably central to our self-conception as agents, there are significant obstacles to the idea that agents are free-willed. When considering the situation of an agent under determinism from the right angle and in the right context, many of us—philosophers and non-philosophers alike—find it hard to understand, at least at first blush, how such an agent could enjoy free will. Reflection on the situation of an agent under indeterminism often has the same effect. In these contexts, the reason why it seems to us that the agent is unfree has nothing to do with any special exculpatory circumstances surrounding her actions. Instead, we have the impression that there are certain very general facts about the causal and nomic relationships in which human actions stand to other factors that preclude the existence of free will. I will call this the "non-existence impression."

Whether the non-existence impression ultimately stands up to philosophical scrutiny is a widely discussed question, but I will set it aside in this paper. I am interested in the prior question of what generates the impression in the first place. What are the considerations that give rise to it (setting aside for the moment the question whether these considerations

are sound)? Presumably, the impression is based on some (perhaps tacit) assumption about the preconditions for free will, and on our impression that these conditions are not satisfied. But what are the assumed requirements for free will, and what are our reasons for thinking that they are not met?

There is no shortage of attempted answers, which often blend into each other. I will sketch two very prominent ones in section 1, before arguing that they do not identify the apparent problem for the idea of free-willed agency (the source of the non-existence impression) in its most general form. The two accounts might do a good job at explaining the non-existence impression for ordinary agents in normal circumstances (and if that is all their proponents are aiming to do, I have no objections), but I will argue in section 2 that they cannot explain why the same impression arises in certain other cases of a more unusual sort. Moreover, I will try to show that we can give a unified explanation of what generates the impression in all the different cases in which it arises. After considering and rejecting two natural accounts in section 3, I will make a new and admittedly speculative proposal in sections 4–6 that relies on the notion of an ultimate explanation of a fact (understood in a special sense that I will explain in section 4: both in ordinary cases and in the extraordinary examples considered in section 2, our impression that the agent is unfree rests on the assumption that free will requires being involved in the right way in the *ultimate explanation* of one's actions, and on the thought that the agent does not satisfy this condition. In fact, given a plausible background assumption, it is impossible for anyone to satisfy this condition, since it is impossible for an action to have an ultimate explanation. This in turn is not a distinctive feature of human actions, but an instance of the general truth that no fact can have an ultimate explanation in the sense that matters.

1. Alternate possibilities and ultimate responsibility

There are different reconstructions of the considerations that generate the non-existence impression, and they differ in the requirements for free will they employ. On one reconstruction, an action counts as free-willed only if it was up to the agent whether she would perform the action—it must have been in the agent's power to do otherwise than she did. Under determinism, so the argument continues, the actions of an agent are determined by facts about the distant past together with facts about the natural laws, and it is not up to the agent what the past or the laws are like. Therefore, under determinism,

(1) Any action is determined by factors that are not up to the agent.

If the action is determined by factors that are not up to the agent, then it is not up to the agent whether she performs the action. Therefore, in a deterministic scenario, it is not up to the agent whether she performs the action, and the action is therefore not performed of the agent's free will. Under indeterminism, an action need not be determined by facts about the distant past together with the laws. However, so the argument goes, the only other

possibility is that the action was (directly or indirectly) the outcome of a chance process. In that case, it was still not up to the agent whether or not she would perform the action, since it is not up the agent what the outcome of a chance process is.[1]

On another reconstruction of the tacit reasoning underlying the non-existence impression, it rests on the assumption that an action is free-willed only if the agent is *ultimately responsible* for the action, in a certain non-moral, causal or explanatory[2] sense—she must be the *ultimate source or origin* of the action. (The proponent of this reconstruction of the non-existence impression would need to explain how to understand the italicized phrases.) Moreover, so the argument runs, under determinism the following is true of any action:

(2) If we follow the causal chain that led to the action in the upstream direction, we will eventually reach a set of causes none of which were brought about by the agent,[3]

such as facts about the agent's genes and upbringing. If an action satisfies condition (2), then the agent is not ultimately responsible for the action. So, under determinism, actions are not free-willed and agents are not responsible for them. Moreover, indeterministic scenarios differ from deterministic ones only inasmuch as the agent's action might be (the direct or indirect) the outcome of a chance process. In that case, the agent is still not ultimately responsible for the action either (for reasons that the proponent of the argument would need to spell out). Therefore, no action is ever free-willed and no agent is responsible for her actions.[4]

I have merely given rough and simplified outlines of the two arguments. There are many different ways of filling in the details. Most parts of the two arguments are controversial. For example, some compatibilists deny that moral responsibility requires that it was in the agent's power to act otherwise, while others accept this requirement but hold that it can be satisfied even if (1) is true. And libertarians might argue (for example by invoking the notion of agent causation) that it can be up to the agent whether to perform a certain action,

---

[1] For arguments along these lines and discussion thereof, see Ginet 1966, van Inwagen 1983, 2002, McKay and Johnson 1996, Speak 2011.

[2] I will use the term "explanation" and related expressions for a metaphysical (not an epistemic) relation: to say that *x* explains *y* is to say that *x* is the reason why *y* obtains, that *x* is responsible for *y*, or that *y* is due to *x*. Causal relationships are paradigmatic instances of explanation in this sense. I will also sometimes speak of "an explanation" of a fact *f* to mean "an account of what explains *f* (in the metaphysical sense described)." The context will always disambiguate.

[3] Here and elsewhere in this paper I am simplifying by pretending that there is only a single chain of explanatory factors that leads to the agent's action. Typically, the explanatory history of the action will instead be a complex structure that keeps branching as you trace it upstream: the action was brought about by the interaction of different causes, each of which was in turn brought about by several different causes, and so forth. Principle (2) can be understood as telling us something along the following lines: if we follow any given branch upstream, we will eventually encounter a cause not brought about by the agent.

[4] For arguments along these lines and discussion thereof, see Strawson 1986, 2002, Kane 1996, 2011.

and that the agent can be ultimately responsible for it, even if the action was the outcome of a chance process.[5] However, my goal is not to decide whether arguments of the two arguments I outlined are ultimately convincing, but only whether they adequately articulate the considerations underlying the non-existence impression. Even if some of the objections against the two arguments are correct, that fact would not undermine the view that the arguments are good reconstructions of the implicit reasoning underlying the non-existence impression. It would merely mean that if they are adequate reconstructions, then the impression rests on a mistake.

I will argue in the next section that there are deterministic scenarios to which the two arguments obviously fail to apply, but which nevertheless give rise to the non-existence impression. That suggests that neither argument provides a completely general explanation of how the impression arises.

2. The curious case of the human *causa sui*

Before stating my key example, I will discuss another case as a warm-up exercise. The two arguments considered in the introduction apply to this example if they apply to any cases.

> *The crime dynasty*. Fred is a career criminal who lives in a deterministic universe whose history is a succession of indistinguishable epochs. He was created by an older crime boss who arrived in a time machine from the previous epoch. Following an ancient tradition, the older crime boss aimed to produce a successor who would resemble him as closely as possible. He created an embryo in his lab, using DNA he had synthesized to match his own, and then raised the resulting child, Fred, in conditions closely controlled to match those of his own childhood. His goal was to ensure that the child would have the same evil character, commit the same heinous deeds, and be equally resolved at the end of his career to continue the crime ring's tradition by traveling to the next epoch to create a replica of himself. And all of this is exactly what the younger Fred ends up doing.

If any agents are deprived of free will and absolved from moral responsibility by the fact that they inhabit a deterministic universe, then surely Fred is such an agent. That is also what the two arguments considered in section 1 have us say. Clearly, all of Fred's actions trace back causally to factors that obtained before he was born, and they were nomically determined by these factors. Moreover, Fred did not bring about these factors nor was it up to him whether they obtained.

Now consider a slightly different story.

[5] For objections along these lines and discussion thereof, see Frankfurt 1969, Haji 2005, Widerker 2002, Fischer 2002, 2011; Ayer 1954; Chisholm 1976, Ekstrom 2000, 2003, Mele 2006, van Inwagen 2002, 2011.

The self-creating time traveler. Twin-Fred is a career criminal who believes that he lives in a universe whose history is a succession of indistinguishable epochs. Moreover, he takes himself to be the scion of an ancient crime dynasty. He believes that an older crime boss arrived in a time machine from the previous epoch and created him (Twin-Fred) as a precise genetic copy of himself, raising him in conditions indistinguishable from those of his (the older crime boss's) own childhood. Having come to the end of his career, Twin-Fred decides to do the same thing. He takes a trip in his time machine to what he takes to be the next epoch, creates an embryo using DNA that he has synthesized to match his own, and raises the resulting child in conditions indistinguishable from those of his own childhood.

At any moment of his life, Twin-Fred is intrinsically qualitatively indistinguishable from the way Fred (in the crime-dynasty example) is at the corresponding moment of *his* life. Moreover, Twin-Fred lives in a universe that is governed by the same deterministic laws as Fred's universe. However, Twin-Fred's beliefs about the universe and about his own life are mistaken. He doesn't live in a universe in which events repeat themselves—there is only one epoch. In fact, the universe began to exist less than a year before his birth. However, from its very beginning the universe contained planets, people, and cities, and was full of spurious traces of a long and rich past. Twin-Fred's time machine does not bring him to the next epoch, as he believes, but instead transports him back to the earliest moment in history. Unbeknownst to him, the baby he creates is himself. Twin-Fred is a *causa sui*, his life forms a causal loop.

In any context in which I have the impression that Fred in the crime-dynasty example is unfree and fails to be responsible for his crimes, I have the same impression about Twin-Fred. It seems clear to me that it would be unfair to absolve Fred from moral responsibility while denying the same benefit to Twin-Fred. There is simply no difference between the characters, intentions and actions of the two agents, or between the causal histories and consequences of their actions, that would justify us in holding Twin-Fred but not Fred responsible. In fact, it seems plausible that any considerations that are relevant to deciding whether Fred is morally responsible should also apply to Twin-Fred.

However, neither of the two arguments considered in section 1 provides a plausible reconstruction of the considerations underlying my impression that Twin-Fred is not responsible for his actions. For, both arguments are blatantly incapable of showing that he is not responsible. Consider as an example Twin-Fred's action of replicating himself (call this action "*D*"). Applied to *D*, (an abbreviated form of) the first argument runs as follows: "(i) There are factors that nomically determine that Twin-Fred performs action *D*, such that it is not up to Twin-Fred whether these factors obtain. Therefore, (ii) it is not up to Twin-Fred whether he performs action *D*. Consequently, (iii) *D* is not a free-willed action and

Twin-Fred is not morally responsible for *D*." The problem with this argument is that there is no obvious reason to accept premise (i). What are the factors supposed to be that aren't up to Twin-Fred and which nomically determine that he does *D*? They cannot be facts about times when Twin-Fred does not yet exist, since there are no such times. Twin-Fred already exists at the earliest moment in the history of the universe, since that is the moment to which his time machine transports him. What about the nomic determinants of *D* that obtain after Twin-Fred has been created in the lab? Presumably, these include the fact that Twin-Fred has certain genes. But why should we think that it is not up to Twin-Fred whether these facts obtain? The relevant facts are, after all, among the effects that Twin-Fred intends to bring about when he decides to do *D*, and they obtain only because Twin-Fred makes the decision to do *D*. Consequently, we would have reasons to deny that the relevant facts are up to him only if we could assume that it is not up to him whether to do *D*, i.e. only if we already had reasons to assume that (ii) is true. But if the only possible reason for accepting (i) rests on the prior acceptance of (ii), then we cannot also use (i) as a premise in an argument for (ii). Therefore, the above argument for (iii) is not cogent.

It should be even clearer that the second argument considered in section 1 cannot be used to show that Twin-Fred isn't responsible for *D*. When applied to *D*, premise (2) tells us this: when we follow the causal chain that led to *D* in the upstream direction, we will eventually reach a set of causes none of which were brought about by Twin-Fred. But that premise is false. Even the causally relevant facts about Twin-Fred's genes and about the environment in which he grew up were (deliberately) brought about by Twin-Fred.

3. Two conceptions of ultimate responsibility

I believe that it is possible to a give a unified explanation of what generates the non-existence impression in the different cases in which this impression arises, including the cases of ordinary agents under determinism and under indeterminism and the example of the self-creating time traveler. Moreover, I believe that the view that explains the impression as resting on the implicit acceptance of the second argument discussed in section 1 is on the right track in one crucial respect: what underlies the non-existence impression is the assumption that free will requires that the agent be in a certain sense ultimately explanatorily responsible for her action, combined with the realization the agent does not meet this condition. However, it is important to understand the notion of ultimate responsibility in the right way. I will argue that on the most suitable way of understanding the concept, we can show that an agent under determinism fails to be ultimately responsible for her actions even when (2) is false. In particular, we can show that the self-creating time traveler Twin-Fred lacks such ultimate responsibility.

For present purposes, let us understand the claim that the agent is explanatorily responsible for an action as (roughly) the claim that the action is explained in the right way

by suitable facts about the agent's intentions or decisions.[6] The crucial question is what we are adding to the claim that the agent is explanatorily responsible for an action when we say that she is *ultimately* (explanatorily) responsible for it. To explain the non-existence impression, we do not need to find necessary and sufficient conditions for ultimate responsibility. It is enough to formulate a *necessary* condition that fails to be satisfied in those examples that give rise to the non-existence impression. In this section, I will consider two very natural attempts to formulate such an account (View #1 and View #2 below), and I will argue that neither of them provides a unified explanation of why we have the non-existence impression in the different cases in which we do. These arguments will prepare the way for a new proposal, which I will describe in next section.

> *View #1*. For an agent to be ultimately responsible for an action, some facts about the agent's intentions or decisions must be at the *beginning* of the chain of explanatory factors that leads to the action.[7,8]

Suppose we combine View #1 with the idea that there are contexts in which we take free will to require ultimate responsibility. We can then explain why it seems to us in such contexts that under determinism agents like you and me are not free. For, the explanatory chain leading to an action of such an agent does not start with any fact about the agent, but instead reaches all the way back to the period before the agent was born. The account can also explain why the self-creating time traveler Twin-Fred appears to be neither free nor responsible. Although Twin-Fred is explanatorily responsible for *D* (his act of self-replication), the explanatory chain that leads to *D* does not *start* with any facts about his intentions or decisions. It does not *start* anywhere at all, since it is a loop with no beginning or end.

However, View #1 cannot explain the non-existence impression in all cases that give rise to it. Consider the following example.

> *Criminal from the get-go*. Ed is a hardened career criminal, who loves his job and his character. He lives in a deterministic universe that sprang into existence complete with planets, cities and people, all going about their business as if they had existed forever. Moreover, Ed was there from the very beginning, and the first

[6] The words "in the right way" and the restriction to "suitable" facts about the agent's intentions and decisions are placeholders for qualifications that are needed to rule out deviant cases. We do not need to worry at present about what counts as an explanatory connection of the right kind or as a suitable fact about the agent's intentions and decisions, as it will not matter for our purposes how exactly we understand these qualifications. (I will drop the qualifications for the sake of simplicity in my discussion below.)

[7] As mentioned in footnote 3, I am simplifying by pretending that there is only a single chain of explanatory factors that leads to the agent's action. The truth is that the explanatory history of the action is likely to be a complex branching structure. View #1 should be understood as telling us that some suitable fact about the agent needs to be at the beginning of at least one of the branches of this structure if the agent is to be ultimately responsible for her action.

[8] For an understanding of ultimate responsibility along these lines, see, e.g., Kane 1996: intro.

> thing he did, at the earliest moment of time, was to cock his gun, ready to rob the local bank. From the first instant, he intended to commit numerous crimes to enrich himself, with no concern for the plight of his victims.

Although the causal chain that led to the bank robbery traces back to (and begins with) Twin-Fred's intention to rob the bank, there is no explanation of the fact that he had this intention. It was not brought about by facts about earlier times (there were no earlier times) nor by facts about later times (there is no backwards causation in Ed's universe). It seems that no one and nothing is responsible for the fact that Ed had this intention. In that respect, Ed's situation is similar to that of an agent under indeterminism whose malevolent intention is the outcome of a chance process. Whatever considerations make us inclined to deny that such an agent is morally responsible should make us say the same about Ed. But this judgment about Ed cannot be explained by View #1. For Ed seems to meet the necessary condition for ultimate responsibility laid down by View #1: the causal chain that led to Ed's robbing the bank started with his intention to do so.

> *View #2*. For an agent to be ultimately responsible for an action, more is required than that the action be explained in the right way by a suitable fact *E* about the agent's intentions or decisions. The agent must also be explanatorily responsible for *E*, in the sense that some suitable fact *E** about her intentions or decisions explains *E* in the right way, and she must also be explanatorily responsible for this further fact *E**, and so forth.[9,10]

This account is very similar to the view that underlies the second argument considered in section 1. We can restate View #2 by saying that an agent is ultimately responsible for an action only if the following is true: when we trace the explanatory chain that led to the action in the upstream direction, not only will we eventually encounter facts about the agent's intentions and decisions, but we will keep encountering such facts no matter how long we keep going in the upstream direction. If we eventually reach a fact that has no explanation or for which the agent is not explanatorily responsible, then the agent is not ultimately responsible for her action. Clearly, this will always happen in the case of ordinary agents in normal circumstances, regardless of whether determinism is true. Consequently, the fact that it seems to us in some contexts that such agents lack free will can be nicely explained by combining View #2 with the thesis that in these contexts we take free will to require ultimate responsibility. But View #2 is of no help in explaining why the self-creating time traveler Twin-Fred appears to lack free will. For, Twin-Fred

[9] Again, I am simplifying (as mentioned in footnote 3). The explanatory history of the action is likely to be a complex branching structure. View #2 should be understood as telling us that the agent is ultimately responsible for her action only if the following is true of at least one branch of this structure: no matter how far upstream you go along that branch, you will keep encountering facts that are explained by intentions and decisions of the agent.

[10] For an understanding of ultimate responsibility along these lines, see, e.g., Strawson 2002.

meets the necessary condition for ultimate responsibility laid down by View #2: when we trace the explanatory chain that led to one of Twin-Fred's crimes in the upstream direction, we will keep going around the explanatory loop, and we will therefore keep encountering facts about Twin-Fred's intentions and decisions.

4. Ultimate explanation and the non-existence impression

Some reflections on the topic of explanation will prepare the way for an alternative view of the conditions for ultimate responsibility. Suppose I ask you what explains a given fact *f*. If you tell me that it is explained by fact *g*, I can follow up with another question: what explains *g*? If you answer this question, you will have given a fuller explanation of *f*—a fuller account of its source and origin—than was provided by your initial answer. (I gain a fuller understanding of how it came about that *f* obtained.) Of course, I can now go on to ask what explains the fact that explains *g*, what explains the fact that explains the fact that explains *g*, etc., at each step requesting an explanation of the fact to which you appealed in answering my previous question. If you keep answering these questions correctly, you will give me an increasingly full account of the source of the initial fact *f*. However, even if this series of why-questions and answers (this "why-regress") goes on forever, your answers may not tell me all I want to know about why *f* obtained. The chain of explanatory factors described by your answers is either circular, forming an explanatory loop, or it is linear, forming an infinite sequence that has no beginning. In either case, I could ask: why does this entire chain of explanatory factors exist? Unless and until this further follow-up question has been answered, I might have the feeling that some question about the source of *f* has not been answered.[11]

To see that this is plausible, consider first the case in which the explanatory chain you describe forms a loop. Lewis considers the example of a causal loop in which a person travels back in time to talk to his younger self and explain to him how to build a time machine. He writes:

> Each event on the loop has a causal explanation, being caused by events elsewhere on the loop. That is not to say that the loop as a whole is caused or explicable. It may not be. Its inexplicability is especially remarkable if it is made up of the sort of causal processes that transmit information. Recall the time traveler who talked to himself. He talked to himself about time travel, and in the course of the conversation his older self told his younger self how to

[11] It is true that your answers explain every part of the causal chain that led to *f* (since every part is explained by a prior part of the same chain). However, it is not at all clear that an account that explains every part of a causal chain amounts to an explanation of why the chain as a whole exists. (This question has been hotly debated, in particular in the context of discussions of the cosmological argument. See, e.g., Hume 1779: part 9, Rowe 1975, Pruss 2003.) I think it is very plausible that an explanation of the parts of the chain does not amount to an explanation of the whole if it invokes facts that are themselves parts of the chain. (See my discussion of the "Externality Constraint" in section 5.) That is why your answers do not seem to explain the existence of the chain as a whole.

> build a time machine. That information was available in no other way. His older self knew how because his younger self had been told and the information had been preserved by the causal processes that constitute recording, storage, and retrieval of memory traces. His younger self knew, after the conversation, because his older self had known and the information had been preserved by the causal processes that constitute telling. But where did the information come from in the first place? Why did the whole affair happen? There is simply no answer.[12]

Lewis's point is not that there *cannot* be an explanation for such a loop but that there *need* not be one, and in particular that explaining each part of the loop by appeal to some other part is not enough to explain the loop as a whole simply by explaining each part of the loop by appeal to some other part. I would add that if there is no explanation of the loop as a whole, then there is a limit to how fully we can explain any given part of the loop (such as the younger time traveler's expertise in time-machine construction). If that doesn't seem obvious, perhaps it can be made clearer by considering a more detailed version of the example. Suppose that I left my office for lunch at noon, after complaining about the fact that no one had figured out how to build a time machine. I return at 1:00pm to find that our colleague Penny now knows how to do it, and I ask you how this came about. You tell me that Penny's older self arrived from the future with blueprints of the time machine and explained the process of building such a contraption to her. When I ask you how the older Penny knew, you tell me that she remembered the conversation she had had with her older self when she was young. All of this tells me less than I had hoped to learn about how the universe came to be one in which Penny knows how to build a time machine at 1:00pm. And that is not because you left out some relevant piece of information. You told me everything there is to know. It is just that there is a limit to how fully we can explain the expertise in time-machine engineering that Penny possesses at 1:00pm.

We confront a similar situation when we replace the explanatory loop with a linear chain of explanatory factors that has no beginning. Suppose that when I ask you how it came about that Penny knew how to build a time machine at 1:00pm, you answer as follows:

> At 1:00pm Penny knew how to build a time machine because she had this knowledge at 12:30pm and her memory of this information was preserved throughout the intervening period. Moreover, at 12:30pm she knew how to build a time machine because she had this knowledge at 12:15pm and her memory of the information was preserved throughout the intervening period. Moreover, at 12:15pm she knew how to build a time machine because she had this knowledge at 12:07:30pm and ….

In some sense, you have told me what explains the fact that Penny knows how to build a time machine at 1:00pm, what explains the fact that explains this fact, and so forth. And yet, there is also a sense in which you have told me less than I would like about how it came about that Penny knows how to build a time machine at 1:00pm. For, your answer

[12] Lewis 1976: 148–9.

keeps me wondering why the whole series of infinitely many consecutive causes exists. Unless and until this further question has been answered, I will have the feeling that there is something about Penny's knowledge at 1:00pm that has been left unexplained. Again, this does not mean that you omitted some relevant piece of information—you might have told me everything there is to know. It might simply be that the fact that Penny knows how to build a time machine at 1:00pm cannot be explained as fully as one might have hoped.

These examples illustrate that we can distinguish two types of follow-up questions to an explanation of a given fact *f*. *Firstly*, given any fact stated in the explanation, we can ask why this fact obtains. *Secondly*, given any endless (in the upstream direction) chain or chains of facts stated in answering previous follow-up questions, we can ask why the collection of these facts as a whole exists. View #2 recognizes the existence of the first kind of follow-up question to a request for explanation. Questions of this kind generate a possibly infinite why-regress, and View #2 is right in telling us that ultimate responsibility requires that the correct answers to all questions in this why-regress state some facts for which the agent is explanatorily responsible. But this necessary condition for ultimate responsibility is not strong enough to explain the non-existence impression in all cases in which it arises. To achieve this, we have to take the second type of follow-up question into consideration as well. Follow-up questions of the two kinds together generate an even more extensive why-regress, one that does not end once we have traced an explanatory chain without beginning that leads to the agent's action. This expanded why-regress gives us a stronger necessary condition for an agent to be ultimately responsible for an action: the correct answer to every question in the extended why-regress needs to state some fact for which the agent is explanatorily responsible.

We can restate this account in slightly different terms. Let an *ultimate explanation* of a fact *f* be an account that correctly answers not only the question "Why does *f* obtain?", but also every follow-up question. Agent *P* is ultimately responsible for her action *A* only if the following condition is satisfied: the ultimate explanation of *A* answers the question "Why did *P* perform *A*?", as well as every follow-up question, partly by appealing to facts about *P*'s intentions or decisions, or to factors partly explained by such facts.

This account imposes a stronger condition on ultimate responsibility than View #2. Consequently, whenever View #2 (combined with the assumption that we take free will to require ultimate responsibility) can explain our impression that an agent lacks free will, the new account can explain the datum as well. The cases that can be explained in this way include those of normal agents in ordinary (deterministic or indeterministic) circumstances, as well as the example of the criminal from the get-go. In addition, the new account can explain the case of the self-creating time traveler, Twin-Fred. Twin-Fred is explanatorily responsible for his act of self-replication, *D*, and for every fact that figures in the circular explanatory chain that led to *D*. But he is not explanatorily responsible for the existence of this circular chain as a whole—the causal loop as a whole has no explanation. That shows

that he is not *ultimately* responsible for any of his actions. Consequently, in any context in which we take ultimate responsibility to be a requirement for free will, we will have the impression that Twin-Fred lacks free will and is not morally responsible for his crimes.[13]

I have suggested that in those contexts in which we have the non-existence impression, we are assuming that ultimate responsibility is a requirement for free will and moral responsibility. But I do not claim that we make this assumption when we form judgments about the blameworthiness of agents in other contexts. It is a well-known fact that the non-existence impression is fleeting. It may have a hold on us in some of our more reflective moments, but it tends to disappear once we move back into a more ordinary frame of mind (even if we continue to think about the same action). One possible explanation for this shift is that we take ultimate responsibility to be a precondition for free will in contexts of the former but not in those of the latter kind.

5. The impossibility of ultimate responsibility

Despite the fact that Twin-Fred is explanatorily responsible for every element of the circular explanatory chain that led to his act of self-replication *D*, he is not *ultimately* responsible for *D*. I argued that this was shown by the fact that he is not explanatorily responsible for the existence of the circular chain (the causal loop) as a whole.

---

[13] The account also predicts, plausibly in my opinion, that the non-existence impression arises in cases where the agent's action is the endpoint of a linear explanatory chain with no beginning, even if the agent is explanatorily responsible for every link in this chain. The following story exemplifies this type of case.

> *The infinite criminal*. Penny is a hardened criminal who lives in a deterministic universe that has existed for an infinite amount of time. Penny herself has always existed and has always worked for the same criminal organization. She loves what she is doing. It is true that she occasionally experiences moments of uncertainty about her life choices. But she never allows herself to remain in that state for long. As soon as she feels a twinge of doubt, she schedules a meeting with the chief recruitment officer of her organization who will eloquently describe the various benefits of the criminal profession, or she slides her fingers through the gold coins in the chest beneath her bed until greed and malice drive all other feelings from her heart. Soon enough, her qualms will dissipate and she will be her old ruthless self. Moreover, she aims to prevent such episodes of weakness, as she calls them, by surrounding herself with unrepentant criminals and avoiding anyone of a different mindset. That is what she has always done, to ensure that she would never reform.

Each of Penny's crimes were caused by some of her earlier intentions and decisions, which were caused in part by various character flaws, which in turn were caused or sustained by her earlier decision to cultivate and strengthen these flaws, and so forth. As in the case of the self-creating time traveler, if we follow the chain of explanatory factors in the upstream direction, we never reach any facts for which Penny is not explanatorily responsible. Nevertheless, on the account sketched in the present section, Penny is not ultimately responsible for any crime she committed. For, there is no answer to the question "What explains the infinite explanatory chain (that led to her crime) as a whole?", and there is therefore no ultimate explanation of the fact that Penny committed the crime. The view therefore predicts that in contexts where we take ultimate responsibility to be a precondition for free will, we have the impression that Penny lacks free will and is not responsible for her crimes.

It is important to note, however, that even if Twin-Fred *were* explanatorily responsible for the loop as a whole, that *still* wouldn't suffice for him to be ultimately responsible for *D*. Another example will make that clear.

> *A crime story with a loopy ending*. Carmen has worked for a crime syndicate for forty years and is very pleased with the way her life has gone. "I would do it all again in exactly the same way," she says. "In fact, I will." Believing that history is running through many epochs during which events repeat themselves, she steps into her time machine and presses the button (at time t_0), intent on traveling to the next epoch. She arrives in place *p* at time t_1. Once again, she works for a crime syndicate for forty years before deciding to take a trip to the next epoch to commit even more crimes. She steps into her machine again and, at time t_2, presses the button. But this time the machine malfunctions, transporting her back to *p* and t_1.

The stage S_1 of Carmen that is located in *p* at t_1 is causally (and psychologically) connected in two different ways to the stage S_2 located at t_2. Firstly, S_2 is the immediate cause of S_1: by pressing the button inside the time machine, S_2 causes S_1 to come into existence in *p* at t_1. But S_1 also causes S_2, indirectly, by way of causing a forty-year-long stretch of intermediate person stages that ends with S_2. S_1 and S_2 are therefore parts of a single causal loop. Moreover, the existence of S_1 in *p* at t_1 is causally over-determined. It is an immediate effect of Carmen's pressing the button inside the time machine at time t_0 and also of her pressing the button at t_2. If you recount the events of Carmen's life in the order in which she experiences them—the order of "personal time," as David Lewis would put it[14]—then S_1 will occur in the story immediately after the button-pressing at t_0 and again immediately after the button-pressing at t_2. After the button-pressing at t_2, the story will keep repeating itself forever.[15]

Consider Carmen's decision just before t_2 to travel to another epoch to commit more crimes. Call that decision "D_2." As in the case of the self-creating time traveler, Twin-Fred, any action Carmen performs after her arrival at t_1 is the result of an explanatory chain that features only facts for which Carmen is explanatorily responsible, and which has no end in the upstream direction (it circles back on itself). However, in this case there is an answer to the question of why the loop as a whole exists: Carmen's decision at t_0 to travel to the next epoch (call that decision "D_0") brought about the existence of the loop. Carmen is therefore explanatorily responsible for the existence of the loop as a whole.

---

[14] Lewis 1976.

[15] In order for there not to be abrupt discontinuities in Carmen's life story at the moments when she travels through time, S_1 must be exactly like S_2, and S_1 must also be exactly like the person stage S_0 located at t_0. So, S_0 and S_2 must be like each other. To ensure that this is so, we can stipulate that Carmen enjoys eternal youth (she does not age), so that the stages S_0 and S_2 do not differ in biological age. We can also make it part of the story that S_0 and S_2 are alike in the contents of their memories. We just need to stipulate that Carmen's life after her arrival at t_1 unfolds in exactly the same way as her life during the forty years before she presses the button at t_0, and that at any given moment she can only remember the most recent forty years of her life.

It should be clear, however, that this gets us no closer to the conclusion that Carmen is ultimately responsible for D_2. For, we can ask a series of further follow-up questions. What explains the fact that Carmen made decision D_0? What explains the fact that explains this fact? … These questions generate another why-regress, and their answers will trace the explanatory chain that led to D_0 in the upstream direction. If this eventually leads us to a fact for which Carmen is not explanatorily responsible, then that establishes that Carmen is not ultimately responsible for D_2. And even if it never leads us to a fact for which she is not explanatorily responsible, that still doesn't show that she is ultimately responsible for D_2. For now we are confronted with a second infinite chain of facts, and we have to go on to ask what explains the existence of this second chain as a whole. We can keep playing this game forever, without ever establishing Carmen's ultimate responsibility for D_2.

In fact, it is easy to prove that it is impossible for an agent to be ultimately responsible for an action, provided we accept the following plausible assumption.

> *Externality Constraint*. A satisfactory answer to the question "What explains this collection of facts as a whole?" must invoke facts that are not themselves part of the relevant collection.

This assumption is supported by its explanatory power. The externality constraint can explain why in the example of the self-creating time traveler, we cannot answer the question "Why did the loop as a whole exist?" by saying that every part of the loop was caused by Twin-Fred's act of self-replication *D*. The reply is true—*D* did indeed cause every part of the loop, including *D* itself—but it does not answer the question on its intended reading. When we are wondering what explains the loop as a whole—why it exists in the first place—we are looking for something outside the loop that explains why the loop exists. Compare the case of Twin-Fred to the crime story with a loopy ending. In that example, we *can* explain the existence of the causal loop as a whole by saying that it was caused by Carmen's decision D_0, since D_0 is not itself part of the loop.

If the externality constraint is correct, then nothing can have an ultimate explanation. Suppose for *reductio* that some fact *f* does have such an explanation, i.e. that there is some account, *A*, that answers not only the question of why *f* obtains, but every follow-up question as well. Now consider the collection of all facts that are invoked in the explanations that are part of *A*, and consider the question of what explains this collection as a whole. Any ultimate explanation of *f* must include an answer to this follow-up question. By the externality constraint, any correct answer to the question will have to invoke facts that are not part of the collection of all facts invoked by *A*. So, the answer to the question cannot be part of *A*. But that conclusion contradicts our assumption that *A* is an ultimate explanation of *f*. This shows that, given the externality constraint, the notion of

ultimate explanation is incoherent, in the sense that the assumption that a fact has an ultimate explanation leads to a contradiction.[16]

If no fact can have an ultimate explanation, then no agent can be ultimately responsible for an action. Consequently, in contexts in which we assume that ultimate responsibility is a precondition for free will, any agent in any possible scenario will seem to us to lack free will.

6. Concluding remarks

Given the externality constraint, the notion of an ultimate explanation is incoherent. However, that does not present a problem for my strategy of using the concept in explaining the non-existence impression. It merely means that the explanation portrays the non-existence impression as resting on our implicit acceptance of an incoherent requirement for free will. That is not obviously incorrect. After all, many philosophers have thought that if the considerations underlying the non-existence impression were correct, they would show that free will is impossible (at least for ungodlike creatures like ourselves) and that the very idea of free-willed agency is incoherent.[17] My account can explain why this is such a natural view to take. The same fact cannot easily be explained if we try to account for the non-existence impression by appealing to one of the two arguments stated in section 1. For, neither one of these arguments establishes the metaphysical impossibility of free will, even for ungodlike human beings. The story of the self-creating time traveler is metaphysically possible and its protagonist is both human and ungodlike, and yet the two arguments fail to demonstrate his lack of freedom.[18]

References

Ayer, Alfred Jules (1954). “Freedom and Necessity,” in his *Philosophical Essays*. New York: St. Martin's Press, 3–20.

[16] The observation that the concept of an ultimate explanation is incoherent in the sense described does not entail that there is in general something wrong with using the notion. Nor does it entail that the question whether a given fact has an ultimate explanation is confused (at least by ordinary standards—perhaps a logically omniscient being would find it confused), or that the impossibility of ultimate explanations is not a substantive metaphysical result. Compare: given widely accepted mathematical principles, the notion of a counterexample to Fermat’s Last Theorem is incoherent, in the sense that the idea that there are counterexamples leads to a contradiction. However, that does not show that the question whether there are counterexamples was (by ordinary standards) a confused or bad question, or that there is something wrong with using the notion of a counterexample to the theorem (for example when raising the question whether the theorem has counterexamples). Even more obviously, it does not entail that Fermat’s Last Theorem is not a substantive result of mathematics.

[17] See, e.g., Strawson 1986. For a more nuanced position, see Smilanski 2000.

[18] For comments and discussion, I am indebted to Julianne Chung, Mark Johnston, Carla Merino-Rajme, Jack Spencer, and the participants of the proseminar that I co-taught with Adam Elga at Princeton during the fall term of 2010.

Chisholm, Roderick 1976. *Person and Object: A Metaphysical Study*, La Salle: Open Court.

Ekstrom, Laura Waddell. 2000. *Free Will: A Philosophical Study*, Boulder: Westview Press.

——, 2003. "Free Will, Chance, and Mystery," *Philosophical Studies* 113: 153–80.
Fischer, John Martin 2002. "Frankfurt-Type Examples and Semi-Compatibilism." In Robert H. Kane, ed., *The Oxford Handbook of Free Will*. Oxford University Press: 281–308.

Fischer, John Martin 2011. "Frankfurt-Type Examples and Semi-Compatibilism: New Work." In Robert H. Kane, ed., *The Oxford Handbook of Free Will*, second edition, Oxford University Press: 243–65.

Frankfurt, Harry 1969. "Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66: 829–39.

Ginet, Carl. 1966. "Might We Have No Choice?" In Keith Lehrer, ed., *Freedom and Determinism*, New York (Random House): 87–104.

Haji, Ishtiyaque 2005. "Frankfurt-Type Examples, Obligation, and Responsibility." *Journal of Ethics* 10: 255–81.

Hume, David 1779. *Dialogues Concerning Natural Religion*. London. Reprinted Indianapolis: Hackett, 1980.

Kane, Robert 1996. *The Significance of Free Will*. Oxford University Press.

——, 2011. "Rethinking Free Will: New Perspectives on an Ancient Problem." In Robert H. Kane, ed., *The Oxford Handbook of Free Will*, second edition, Oxford University Press: 381–406.

Lewis, D. 1976. "The Paradoxes of Time Travel." *American Philosophical Quarterly* 13: 145–152.

McKay, Thomas and David Johnson. 1996. "A Reconsideration of An Argument Against Compatibilism." *Philosophical Topics* 24: 113–22.

Mele, Alfred 2006. *Free Will and Luck*, New York: Oxford University Press.

Nagel, Thomas 1976. "Moral Luck." *Aristotelian Society Supplementary Volume* 50: 137–51.

Pruss, Alexander. 2003. “The Hume-Edwards Principle and the Cosmological Argument.” In Richard Gale and Alexander Pruss, eds., *The Existence of God*, Burlington, VT (Ashgate): 347–63.

Rowe, William 1975. *The Cosmological Argument*. Princeton: Princeton University Press.

Smilansky, Saul. 2000. *Free Will and Illusion*. Oxford (Oxford University Press).

Speak, Daniel. 2011. “The Consequence Argument Revisited.” In Robert H. Kane, ed., *The Oxford Handbook of Free Will*. Oxford University Press: 115–30.

Strawson, Galen. 1986. *Freedom and Belief*. Oxford: Clarendon Press.

——, 2002. “The Bounds of Freedom.” In Robert H. Kane, ed., *The Oxford Handbook of Free Will*. Oxford University Press: 441–60.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

——, 2002. “Free Will Remains a Mystery.” In Robert H. Kane, ed., *The Oxford Handbook of Free Will*. Oxford University Press: 158–80

——, 2011. “A Promising Argument.” In Robert H. Kane, ed., *The Oxford Handbook of Free Will*, second edition, Oxford University Press: 475–83.

Widerker, David 2002. “Responsibility and Frankfurt-Type Examples.” In Robert H. Kane, ed., *The Oxford Handbook of Free Will*. Oxford University Press: 323–36.

Williams, Bertrand 1976. “Moral Luck.” *Aristotelian Society Supplementary Volume* 50: 115–35.