Feeling Free and Moral Relevance

Conor O'Brien

4,118 words

1. Overview

Invariably, the more terms involved in a definition or claim, the wider and more unmanageable the surface of counterargument becomes. I therefore take minimalism as a virtue in my claims. I shall begin with a fairly simple thesis about moral responsibility and attempt to elaborate upon the terms I use.

The Free Feeling Claim (FF): An agent is *morally responsible* for an action and its *reasonably foreseeable consequences* if and only if that agent *feels free* in doing that action.

In the next section, I shall explore the exact meanings of the highlighted terms and their implications. Then, I will supplement FF with the following claim:

The Moral Relevance Claim (MR): FF provides a sufficient account for all the relevant operations we can perform upon moral truth claims.

Notice that I do not make any claims about whether or not the world, or any particular aspect of it, is deterministic. In that sense, I am a compatibilist: Even if the world or aspects of it are deterministic, the feeling of freedom still persists, and I would still attest to the definition of moral responsibility under FF.

2. Clarifying Terms in the Feeling Free Claim

There are two primary terms in FF as stated in §1 I would like to clarify and discuss:

1. *Feeling free.* The nature of feeling free is complex, and it may not accurately describe reality in some cases, such as those where one wrongly believes they are free.
2. *Reasonably foreseeable consequences.* To what extent can we reasonably expect an agent to foresee consequences?

2.1. On Feeling Free

In stipulating that an agent must feel free in doing their action, we introduce a considerable amount of imprecision into our definition. Feelings tend to be subjective, and one might think, form no strong basis for a definition. With sufficient analysis, however, I believe I can classify most of the feelings of action freedom.

We can vary whether or not someone truly does feel free as well as whether or not they in actuality are free, giving us four broad, possible cases. I shall create terms for agents of each

of these categories. The agents who are *rationals* have alignment between their belief about their freedom and the actuality of their freedom; there are *freedom rationals*, who are free and feel free, and *determined rationals*, who are not free and do not feel free. The agents who are *delusionals* are those where their belief and the actuality of their freedom are at odd; there are *freedom delusionals* who believe they are free but are wrong in so believing and *determined delusionals* who believe they are not free but are likewise wrong.

The cases of freedom rationals and determined rationals are in accordance with FF, as they are the true-true and false-false cases of the claim's main biconditional. The functional feature in each of these cases which makes their truths evident is the truth value of their freedom.[1] To elaborate, we would naturally hold freedom rationals morally responsible for their actions, and contrariwise not hold determined rationals morally responsible.

The two significant cases that remain were those previously exposited: Freedom delusionals and determined delusionals. Before discussing these two cases, I must note the issue of *extent*. Whereas the categories I have drawn here are fairly discrete (is free, does not feel free), the truth is often much fuzzier (is partially free, does not feel entirely free); the extent to which one feels or is free is often not a binary classification. Although I will continue arguing in the next sections as if the categories I have outlined were truly discrete, but one may easily substitute in the appropriate fuzzy logic operators for suitably fuzzy truths about the freedom and feelings involved.[2] This entails that the fuzzier the truths, the fuzzier the outcome, which means FF, when applied to fuzzier truths, is less immediately applicable; the same could be said about most arguments leveraging fuzzy logic.

2.1.1. Freedom Delusionals

There are two broad categories that I will further separate freedom delusionals into. First, we have the *exceptional freedom delusionals*, who are deluded into falsely believing they are free due to some exceptional properties, such as having one's brain chemistry unsurpassably altered by addiction or illness. Second, we have the class of *quotidian freedom delusionals*, people who falsely believe they are free for some other, more mundane reason than an exceptional property.

---

[1] I presume we can hold those who are in actuality free morally responsible, so too we cannot hold those who are in actuality not free morally responsible; demonstrating this is beyond the scope of this paper.

[2] Dissecting the precise nature of the boundary between significant and insignificant differences would probably make for some interesting semantic arguments. However, this need not be clarified here, since all that I need functionally from significant is that we do not confuse minor incongruities, such as where the agent feels, say, 90% free, but in actuality is only 85% free. Hence, we want some fuzziness in our classification as opposed to a strict mathematical inequality, but the specific degree or boundary is difficult to parse and may even vary between cases; such discourse is beyond the scope of this paper.

Examples of either subclass of freedom delusionals seem to demonstrate the *un*reliability of freedom feelings and consequently undermine FF's very premise. So, in defending FF, I must either show these examples to be not fatal to FF or show them to truly not be examples.

I will begin by consider exceptional freedom delusionals and a motivating example. Consider the shopping addict who might truly believe they are free in their choice to binge shop every Friday evening. In reality, they are compelled to binge shop by some underlying mechanics of how their brain is wired. Consider alternatively the chronically deceived, or those who suffer from a debilitating mental illness. While as harrowing some of these cases may be, it is clear that FF cannot adequately handle them, since their feelings are unreliable due to some atypical mental configurations. This is a shortcoming-turned-feature of FF that I am willing to accept. Perhaps future work can augment FF to apply to exceptional freedom delusionals more generally.

As for quotidian freedom delusionals, I believe this set is empty. Unlike with exceptional freedom delusionals, I am unwilling to accept examples of this kind, as it would underwrite the applicability of this argument for the broader population beyond exceptions as mentioned above[3]. I will examine three classes which could possibly constitute quotidian freedom delusionals and demonstrate why I think they are not truly such.

First, it may seem the class of people who have false beliefs about their actions could be quotidian freedom delusionals. Consider the example of an avid student who has a lucky pencil they use on their most important exams. They seem to believe, on some level, that by using this pencil, they will do better on their exam. Supposing they do well on their exam, they might have the belief that their choice in using the lucky pencil is substantially responsible for the positive outcome. In some sense, we could say this is something they freely believe, despite it being false. However, this freely chosen false belief is not one about the nature of choice, but rather, the impact and interpretation of choices, a body of discussion unrelated to the topic at hand.

---

[3] Although, one could also argue that my notions of "exceptional" and "quotidian" are thinly veiled and potentially misguided normative claims akin to labels of "abnormal" and "normal". Maybe the reality of the human brain or psyche is that we are ultimately incredibly unreliable, and that FF is too naïve a claim to adequate account for the inconsistency in the human brain. Maybe, a large swath of the population could be misguided as to their feelings of freedom and may be more similar to willing addicts deceived about the true reality of their actions. Even still, I would argue that this is unlikely given how widespread notions of conventional moral responsibility are, and I have cautious optimism in people's ability to ascertain how free they feel in acting.

Next, I will consider a study done by Johansson et al.[4] In exploring a phenomenon they term *choice blindness*, participants were asked to choose between two pictures of female faces on what they found attractive. Using slight-of-hand techniques, the participants were then given the picture opposite the one they choose and asked to justify their reasons for picking it. Most participants failed to notice the swap occurring, but nonetheless offered justifications for why they choose it.

The choice blindness phenomenon points out that the human psyche is fantastic at post-hoc rationalization. It also seems to undermine the idea of using the feeling of freedom as a determining factor in moral responsibility, since the subjects in this scenario would attest to freely choosing the portrait given to them despite having chosen the opposite. Yet, what matters in theory is not what the agent would attest to, but the nature of the actual choice the agent made. At the moment of choice, the agent felt free and actually chose freely; it was then the experimenter who intervened upon this decision, presented the opposite choice, and procured false rationalizations from the participants. The flaw here is not a flaw in FF, but in the human psyche. Thus, in the choice blindness study, we cannot consider the participants true freedom delusionals, but rather as freedom rationals who were thereafter misled both by researcher and their own psyche[5].

Last, I will examine a theoretical example. Suppose there is some evil demon, or a sufficiently capable neuroscientist, who influences or even directly causes decisions I nonetheless feel free in executing. This seems the clearest counterexample to FF, featuring the truest and purest feelings of freedom contrasted with virtually no actual freedom involved. However, my counterargument is rooted in practicality. Are there such demons or neuroscientists? I contend they exist only in the theoretical realm, and the burden of proof lies on establishing their existence. I will accept fully that, if such demons or neuroscientists exist on the large scale, FF is an unreliable if not outright incorrect claim. If instead they exist in specific, niche cases, then I would probably classify them more as exceptional freedom delusionals and likewise accept them as simply not falling under FF. Of course, if they do not exist at all, then FF suffers naught.

Generally, in considering these cases, it seems most quotidian freedom delusional contenders could be explained away in terms of freedom rational actions with some posterior or

---

[4] Johansson et al. (2014). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It! *Journal of Behavioral Decision Making*, *27*(3), 281–289. https://doi.org/10.1002/bdm.1807

[5] In practice, the shortcomings of the human psyche to handle cases like this may have widespread implications for the execution of justice and testimony. This practical worry, however, applies broadly to most philosophy about the nature of choice.

exterior delusion separate to the feeling relevant to the action itself. Thus, I believe the set of quotidian freedom delusionals is an empty one, as I imagine most potential examples, similarly to the examples I have considered, can be more correctly categorized differently, either as true freedom rationals in the moment, or as exceptional freedom delusionals, whose existence I tolerate under FF.

There now remain two concerns about freedom delusionals in general that I can identify. First, if determinism is true, in part or in whole, then the set of all who feel free is indeed the set of freedom delusionals in deterministic cases. Second, choice blindness raises demonstrates the human psyche *in general* is, at times, unreliable, which would call into question the more general usage of feeling free in FF. I admit both of these freely and will argue later, through the Moral Relevance Claim, that neither is ultimately a fatal objection.

2.1.2. Determined Delusionals

Determined delusionals are people who believe they are not free but in actuality are. The issue such people pose for FF is that it seems we are unable to ascertain whether or not they are ever morally responsible; indeed, the biconditional in FF would say they are *not* morally responsible for their actions. That people may simply bypass any moral responsibility under this claim by simply believing they are not free in their actions is concerning for its validity.

As I will show in §2.2., there are certain Rational Imperatives we impose on agents to judge their moral responsibility. If we could rationally expect determined delusionals to be able to see their own delusion, then this class of agents poses no threat to FF. However, if it turns that it is rational to believe some or all actions are determined even in an overwhelmingly libertarian appearing world, perhaps via a stance along Plantinga's reformed epistemology[6], then FF would seem to fail to apply to determined delusionals who are, almost paradoxically, rational in their delusions. To this end, I have no solid counterargument, save that it seems unlikely such a defense of determinism in such a world can be made.

2.2. On Reasonably Foreseeable Consequences

Before I begin my discussion here, although considering what delineates action is of notable philosophical interest, it is beyond the scope of this paper to argue in great detail. I am primarily inspired by Davidson's account of action and reasons[7], which, in short, states that an

---

[6] https://iep.utm.edu/ref-epis/, e.g.

[7] As expressed between Davidson's "Actions, Reasons, and Causes" (1963) and "Agency" (1971). While there is much debate on Davidson, all that my account requires is a suitably consistent theory of intentions, actions, and agency, such as Davidson provides; the minutiae of the issues resultant from such a theory, such as akratic (weak-

agent performs an action for a primary reason, under some description of that action. For example, I may intentionally turn on a light switch to illuminate a room, but even if there is a burglar in that room, and in so turning on the light switch I alert him to my presence, this was not an intentional action of mine.

Now, FF seeks to make an agent morally responsible both for the action they do, as well as that action's *reasonably foreseeable consequences*. I separate an agent's actions from that action's consequences because of how differently the agent perceives these things. While the agent might have high confidence in how the action they plan to take occurs, the further in the future the agent tries to project that action's consequences, the less confident the agent often is in their appraisal.

To illustrate this confidence, consider an advanced billiards player. She knows many tactics and strategies for manipulating the billiard balls with the cue in ways favorable to her game-playing. She wishes to strike the cue ball with enough spin such that it arcs around her opponent's striped ball and hits her own solid ball into the far-right pocket. As an advanced billiards player, she has a high degree of confidence in her ability to strike the cue ball with her cue to produce a certain amount of spin. She is also confident that the ball will indeed arc around her opponent's ball and head in the direction of her own ball. However, as she is not an expert, she has less confidence (but still a fair amount) that the cue ball will then hit her own ball in the way she desires it—into the far-right pocket. Furthermore, she is even less confident that the cue ball will end up in a convenient enough place for her next shot, supposing she even sinks a ball in on her current shot. And so on, she is decreasingly certain about the outcome of the game as the tree of potential outcomes explodes with possibilities.

To some extent, when we plan to take an action, we are like the billiards player, assessing what might happen when we start action. This process of assessing the situation and acting accordingly may happen incredibly dynamically, such as in many athletic sports, or it may happen quite statically. It may happen with incredible forethought, such as in chess, or in the spur of a moment, such as reacting to someone charging towards you brandishing a knife.

The amount of planning we do helps reveal the future, and this explains the foreseeability term in FF. Pulling the trigger to a gun aimed at someone almost necessarily entails them being shot, and (if aimed suitably) usually entails them dying without further assistance. They who pulled the trigger is therefore responsible for the death of the they who was shot, as that death is part of the foreseeable consequences resulting from pulling the trigger.

---

willed) actions, have been discussed at length in philosophical discourse, and are not the primary focus for this paper.

However, foreseeability alone is insufficient for determining moral responsibility of consequential actions. The agent must have been able to *reasonably* foresee the particular consequence to be morally responsible for it. For example, we would not hold the agent who, upon flipping a light switch which was wired to an explosive device, morally responsible for detonating said device, assuming they had no idea such a device was there. While the agent could have, with sufficient imagination and time, foreseen that an explosive would be wired to the light switch, this would number among trillions of potential fictitious scenarios, none of which we would reasonably expect the agent to consider when performing what is otherwise such a mundane action as flipping a light switch. Few reasonable agents, if any, would foresee such an explosive, despite it still being a possible foreseeable consequence.

In cases such as the light switch explosive, it is overwhelmingly evident that some situations cannot be reasonably foreseen. Yet, in many other cases, it is much less clear. I claim we can reasonably expect an agent, in considering whether or not to do a particular action, to foresee a particular consequent **C** if and only if the agent meets *at least one* of these conditions (which I shall term **The Reasonability Imperatives**, or RI):

R_1. They should have had sufficient reason to believe **C** could occur.
R_2. They should have had sufficient opportunity to consider whether or not **C** could occur.
R_3. They should have let themselves be in a situation where R_1 and R_2 were true.

If an agent does not satisfy any of these three imperatives with regards to **C**, then it follows that we could not have reasonably expected them to have foreseen **C**. By FF, the agent is therefore not responsible for **C**. At first, R_3 may seem tautological, but it serves the purpose of embedding the concept of action tracing into RI. R_1 and R_2 alone do not adequately match our reasonable expectations in cases where some prior decision prevents R_1 and R_2 from occurring, say, through voluntary drunkenness.

To illustrate how these imperatives work out, consider the following case. Suppose Sam is participating, unbeknownst to them, in a cruel experiment. They wait in a room in front of a dashboard, upon which are two buttons, their labels obscured from Sam's view. Sam is told they must press one of the two buttons within three seconds of their labels being revealed, or else a series of explosives will go off in a nearby building, killing everyone in it. After minutes of anxiety-packed waiting, the experimenter reveals the labels to Sam. The first button is labeled "Detonate the explosive," and the second, "Disarm the explosive." Sam, in their panicked state, feels pressured to immediately press one of the buttons disregarding the labeling, and ends up hitting the first button. The explosive goes off, killing everyone in that building.

We can go through each of the RI, one by one, and examine whether or not we should blame Sam for their recklessness in this situation, recalling that we should blame Sam if they satisfy any of these imperatives. First, consider R_1 only holds if we would reasonably expect

Sam to have read the labels and contextually believe they would have the effects as labeled. In defense of Sam, one could argue they do not satisfy R_1, as they acted immediately and did not have enough time to have read the labels, and therefore we could not have expected them in such a state to do so.

Consider then R_2: We should have expected an agent in Sam's situation to take as full advantage as possible of the allocated three seconds. Once again in Sam's defense, one could argue Sam also does not satisfy R_2 and say that they were so panicked that they could not have acted otherwise, thereby removing from them any imperative expectation of their opportunity for considering the outcomes of their actions.

Supposing one convincingly argues for Sam's innocence of R_1 and R_2, whether or not we blame Sam in these circumstances depends on whether or not we could have reasonably expected them to have a calmer head in that situation, for example, as per R_3. Although this answer may differ depending on your interpretation on the scenario, or factors not explicitly stated in the scenario, this is the most reasonable approach we can take regarding the question of Sam's moral blameworthiness.

Different situations might give different parameters for the RI, and thus different outcomes. Varying R_1, if instead the labels had no text on them, nor any identifying features, yet the same properties of detonating and disarming, we would not blame Sam for pressing the wrong button, as they had no reasonable way of determining which button to press, or even that there would be a consequence for pressing either. Varying R_2, if instead Sam had the entire day to choose which button to press, we would be more likely to say Sam definitely should have taken advantage of the bountiful time present to them, and comfortably blame them for prematurely pressing the wrong button. If Sam had no time at all, we would comfortably excuse Sam from any blame, as no rational agent could have parsed the labels instantaneously. Varying R_3, if instead Sam had suffered from, say, dyslexia, by no fault of their own, and could not read the text on the labels within the allotted three seconds, we would likewise not likely blame them.

I believe I have shown RI provide a powerful tool for evaluate most situations and allow us to assign moral responsibility and a fair and just way.

3. Defending the Moral Relevance Claim

Let me requote the Moral Relevance Claim as it was stated in §1:

> **The Moral Relevance Claim (MR):** FF provides a sufficient account for all the relevant operations we can perform upon moral truth claims.

By relevant operations, I mean those practical applications that we may derive from any philosophical system about moral responsibility, such as determining guilt and assigning blame, praise, and punishment. My claim, then, says that FF should work in practice as a fairly

consistent system for all these different moral considerations, notwithstanding the various provisions I have identified previously.

As mentioned earlier in §2.1.1, psychological concepts such as choice blindness call into question the reliability of the human psyche. While there are a wealth of experiments corroborating the many particular shortcomings of the human psyche, I generally doubt the widespread ramifications these experiments have in day-to-day life. Suppose we apply choice blindness to a case where we must make a moral decision. Since it is a phenomenon affecting how we attest to our beliefs in the moment, it only matters insofar as it directly affects our ability to recount what has happened. This may make it harder to get at the truth of what happened in a courtroom setting, but the fact remains that in cases like these, the truth was that the agent acted freely and thought they were free, and that only their recollection of this state is muddled or even inaccessible.

Additionally, MR is an important claim because it frees FF from dealing with some of the potential reality of determinism. In a way, it is similar to a pragmatic view of moral responsibility of blaming people as a justification for, say, imprisonment, because it is a convenient way to prevent further harm from the agent, as well as discourage future harm from other agents. Importantly, however, the justifying measures we use are not based on a metric of utility, but of the inner workings of FF described previously. And, since FF is based on reasonable expectations and agential feelings of freedom, it is a significantly more attractive and humane system that avoids the cold utilitarian feel of pragmatism.