UNBOUND: FREE WILL FOR MORAL RESPONSIBILITY SKEPTICS

by

Hoi Yee Chan

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF PHILOSOPHY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2022

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Hoi Yee Chan
titled:

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Carolina Sartorio
Carolina Sartorio Date: Jul 25, 2022

Michael McKenna
Michael McKenna Date: Jul 25, 2022

Keith Lehrer
Keith Lehrer Date: Jul 25, 2022

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Carolina Sartorio
Carolina Sartorio Date: Jul 25, 2022
Dissertation Committee Chair
Philosophy

Acknowledgements

I would like to thank each of my committee members—Carolina Sartorio, Michael McKenna, and Keith Lehrer, for their guidance and support. I am particularly indebted to Keith Lehrer, without whose continuous encouragement I might not have finished this dissertation. I would also like to thank Shaun Nichols for inspiring this dissertation and supervising it until he left Arizona. This dissertation has also been improved by feedback from Elijah Millgram, for which I am grateful.

For emotional and other forms of support, I would like to thank my family, Nirmali Fenn, Barbara Sakowicz, Satya Zawaski, Julie Braunstein, Isabel Fernandez-Sirgo, Gloria Woo, Bill Stubler, Tyler Millhouse, Lucia Schwarz, my friends from the University of Arizona Hong Kong Student Association, and my best mate Antonio Cafiero.

Dedication

Dedicated to my maternal grandmother, Yuk Chu WONG, who will never read this dissertation

Table of Contents

Abstract

Many contemporary discussions of free will share the assumption that philosophical questions related to free will are important only for reasons related to moral responsibility. In this dissertation, I explore an alternative explanation for why we may care about free will—liberty appears empty without it. This dissertation is composed of three independent chapters, which respectively address the following questions:

1. What kind of free will can we hope for?
2. What is a free agent like?
3. Why should we express our resentment if moral responsibility does not exist?

Introduction

Suppose we lack moral responsibility in a basic desert sense because determinism is true, or because *causa sui* is impossible, as Galen Strawson argues. Does it entail that we also lack free will? The answer to the question is yes, if we, following a predominant approach to addressing the free will debate in the literature, define "free will" as an agential characteristic in virtue of which one is morally responsible. In this dissertation, however, I will propose a way to think about free will that allows us to affirm free will while denying the possibility of moral responsibility in a basic desert sense.

In my estimation, the primary task in addressing the free will debate is to settle on the meaning of "free will". A traditional approach to undertaking this task is conceptual analysis, which aims to describe the conditions under which the pre-theoretical "folk" concept of free will apply. There are two shortcomings to this approach. First, there may not be a coherent, unified folk concept of free will; the folk concept likely contains conflicting elements[1]. Second, the folk concept of free will likely contains libertarian elements[2], which many philosophers believe are erroneous. My alternative, revisionist approach to settling on the meaning of "free will" is conceptual engineering. By examining why we may want to have a concept of free will, we can decide the conditions under which the concept applies. In Chapter One, I argue that a primary reason for why we may want to endorse a compatibilist concept of free will, and thus affirm the existence of free will, concerns motivation. It is a common observation that free will skepticism erodes

[1] See Chan, Deutsch and Nichols (2016).
[2] See Knobe and Nichols (2007).

motivation. This observation is often explained in terms of moral responsibility. As an alternative diagnosis of this observation, I suggest that free will skepticism erodes motivation because it undermines the perceived meaningfulness of liberty. I propose a compatibilist concept of free will that captures the agential features that enable liberty to be exercised meaningfully. According to my account, free will occurs on a spectrum. On the less ambitious end of the spectrum, free will is exemplified by reasons-responsive actions; on the more ambitious end of the spectrum, free will is exemplified by actions that are not only reasons-responsive, but also original and novel. I argue that someone who strives to realize the more ambitious end of the free will spectrum maximally deserves liberty by exercising it meaningfully, even though they may not deserve praise and blame for reasons related to skepticism about moral responsibility.

In Chapter Two, I explore the question of what a free agent is like if free will requires practical rationality. Against the widely shared view that free will is compromised when luck is an element of one's decision-making process, I argue for the compatibility between free will and agent-internal luck by examining the dominant accounts of practical reasoning. Some compatibilists—notably Mele (2006) and McKenna (draft)—think that while agent-internal luck is compatible with free will, it is nonetheless an undesirable feature that a free, rational agent would strive to eliminate. Call this view *restrictive compatibilism about agent-internal luck*. Against this view, I argue that optimal rational control is compatible with agent-internal luck. Moreover, I argue that more practically rational agents tend to exhibit more agent-internal luck in their decision-making processes.

A question faced by all moral responsibility skeptics concerns whether resentment should ever be expressed. Some philosophers (call them the *revisionists*) say no, and others

(call them the *preservationists*) say yes. Those who favor expressing resentment often justify their position by appeal to the moral benefits of doing so. Against those preservationists, in Chapter Three, I argue that the moral benefits of expressing resentment are quite limited, citing relevant empirical evidence. Despite the limited moral benefits of expressing resentment, I argue against the revisionists that we should nonetheless express resentment sometimes for the sake of emotional intimacy, which is an essential part of a flourishing human life. I propose a version of preservationism—*restricted preservationism*, which recommends limiting the expression of resentment to contexts where emotional intimacy is prized. My position can be considered an effort to reconciliate the two extreme positions of revisionism and standard preservationism, which can be endorsed without assuming that people deserve blame in a basic sense.

Chapter One

Freedom beyond Rationality

1. Introduction

In this chapter, I propose a compatibilist concept of free will that even a moral responsibility skeptic can endorse. Even if moral responsibility does not exist, we still have good reasons to affirm free will, for denial of free will tends to erode motivation. In the next section, I will examine why denial of free will tends to erode motivation. My diagnosis suggests that denial of free will erodes motivation not because it results in denial of moral responsibility, but because liberty appears empty without free will. Then, in section 3, I present my proposed account of free will, which is an extension of traditional reasons-responsiveness accounts. My account can be seen as providing the ultimacy condition of free will, which I argue is preferable to McKenna's (2008) rival accounts in section 4. After that, in section 5, I explain why adopting my account would enable one to deserve not only an irreplaceable position in others' conception of utopia, but also liberty.

2. Engineering a useful concept of free will

It is sometimes said that the free will debate is verbal. Chalmers (2011), for instance, argues that since compatibilists and incompatiblists tend to use the term "free will" to mean different things, many disagreements between them will be dissolved once they clarify their theses without using the term. An assumption behind Chalmer's argument is that the term "free will" can be legitimately used to refer to either a compatibilist or an incompatibilist notion. This assumption has been confirmed empirically by Nichols, Pinillos

and Mallon (2016), who provide experimental evidence that natural kind terms like "free will" are systematically ambiguous. Specifically, the reference of a token term is fixed by a descriptivist convention in some contexts and by a causal-historical convention in other contexts. Applying this experimental finding to the free will debate, Nichols (2015) notes that the term "free will" can be fixed by either a descriptivist convention or a causal-historical convention, allowing us to either affirm or deny the existence of free will legitimately. So, what should we do? Nichols further argues that whether we should affirm or deny free will in a particular context depends on our practical interests. In particular, Nichols argues that denial of free will is likely to interfere with our productivity. By contrast, adopting a compatibilist concept of free will and thus affirming our status as free agents is conducive to our productivity. Our primary interest in adopting a compatibilist concept of free will, then, concerns motivation.

Nichols's suggestion that adopting a compatibilist concept of free will is beneficial for motivation is plausible. After all, it is a common observation that denial of free will can erode motivation. However, Nichols doesn't explain why that is the case, nor does he specify which compatibilist concept of free will we should adopt. In the following, I will take up Nichols's unfinished business by engineering a compatibilist concept of free will that would be beneficial for motivation when adopted.

Why does denying free will erode motivation? There are two possibilities. The first possibility concerns moral responsibility. It is often suggested that in denying free will, one also denies moral responsibility. Since moral responsibility is a source of motivation, denial of free will and moral responsibility may undermine one's motivation. If denial of free will erodes motivation because it denies moral responsibility, the remedy for it would involve

adopting a concept of free will that allows us to affirm moral responsibility. Indeed, revisionists (such as Vargas, 2007) often develop concepts of free will that are specifically designed to justify practices of attribution of moral responsibility. Compatibilists and incompatibilist likewise often take an account of free will to be correct to the extent it captures the relevant agential features that are required for one to be an apt target of moral responsibility[3]. In my estimation, however, moral responsibility isn't the primary reason that denial of free will erodes motivation. There are three considerations. First, attribution of moral responsibility can be a source of motivation even if one denies free will and moral responsibility in a basic desert sense. Whether or not I am a free will skeptic and believe in moral responsibility in a basic desert sense, other people, who are usually not skeptical about moral responsibility, will continue to praise and blame me. Suppose my boss is a compatibilist and blames me; even if I believe that I don't deserve to be blamed, his blaming me would still likely cause me distress and thus affect my motivational system. On the other hand, suppose my grandmother praises me; even if I believe that I don't deserve to be praised, her praising me would nevertheless put me in a positive mood and motivate me to earn more undeserved praise from her. Therefore, denial of free will at a personal level (rather than at a collective level) has minimal effects on our motivation to respond to praise and blame.

Second, for a mature person, attribution of moral responsibility is at best a secondary source of motivation. While we have good reasons to manage our social relationships, and respond to others' moral evaluations of us whether we find them justified or not, our primary moral motivation should not be praise and blame from other

[3] For example, see McKenna and Pereboom (2016).

people. The reason is that in actual practice, praise and blame only indicate praiseworthiness and blameworthiness with limited reliability[4]. Unless one is in a culturally homogeneous environment, one is also likely to encounter conflicting reactions from different people for the same action. A mature person should therefore readily follow their moral compass, independent of praise and blame from other people.

Third, moral responsibility has been overemphasized in contemporary discussions of free will. Most of our actions, such as how we dress or whom we spend time with, are unlikely to provoke praise or blame from other people. If denial of free will only erodes motivation for reasons related to moral responsibility, denial of free will would not affect our motivation for most actions. But people who take the thesis of free will skepticism

[4] To see that moral praise and blame do not track praiseworthiness and blameworthiness, consider cases in which praise is present when it isn't called for, praise is absent when it is called for, blame is present when it isn't called for, and blame is absent when it is called for. I will make my case by borrowing Scanlon's (2013) account of blame, which essentially involves modification of one's attitude towards the person being blamed in a negative way, such as "decreased readiness to enter into special relations such as friendship with the person"(p.105), and "decreased willingness to help the person with his projects"(p.106), "decreased tendency to take pleasure in things going well for the person and to feel sad or regretful when they do not" (p.106). Extending Scanlon's account to praise, praise would involve modification of one's attitude towards the person being praised in an analogously positive way. I prefer Scanlon's account over others because it captures the sort of social rewards and punishments that we actually care about, and to which we are likely to respond. For cases in which praise is present when it isn't called for, consider all the historical cases where aggressors are celebrated as national heroes, such as Hitler and Trump. Although those cases are relatively rare, they highlight the fact that people are ready to praise those who are not praiseworthy. Or consider how often toxic masculinity is praised in the United States; praise sometimes arises from bad norms rather than objective moral reasons. For cases where praise is absent when it is called for, consider cases where cultural norms deem praiseworthy actions as not praiseworthy. For instance, when slavery was considered legitimate in the United States, helping a slave escape from their owner, which is praiseworthy, would result in blame rather than praise. Or consider countries like Afghanistan, where a woman who pursues education would receive blame rather than praise. Our culture isn't immune from such biases. Praise can also be elicited when someone has the appearance of virtue but is not in fact virtuous. Aung San Suu Kyi, for instance, received praise internationally for a long time before her corruption came to light. The cases of slavery in United States and of women in Afghanistan also highlight how blame can be present when it isn't called for. Other examples involve racism, sexism and discrimination against the LGBTQ+ community. Through no flaw of one's own, one may receive blame simply because of one's ethnic background, gender, or sexual preference. In some extreme cases, members of those minority groups are targets of aggression, such as mass shooting, which is a radical manifestation of blame on Scanlon's account. Finally, for cases where blame is absent when it's called for, we can consider cases where people cover up their wrongdoing, or cases where wrongdoing is considered legitimate in a culture, such as slavery or colonization, in the past.

seriously sometimes experience am existential crisis[5]. That is, denial of free will may result in global erosion of motivation, including for actions that are unlikely to result in any praise or blame. We should therefore seek an alternative explanation as to why denial of free will erodes motivation.

There is an alternative possibility as to why denial of free will erodes motivation, which to my knowledge has not been discussed in the literature. The possibility is that denial of free will erodes motivation because it undermines the perceived meaningfulness of liberty, or external freedom. Liberty is often considered to be one of our aspirations with the most intrinsic value; it is also one of the things that we are most intrinsically motivated to pursue. Some may even say that liberty is what we are *most* motivated to pursue. However, if we deny free will, the appeal of liberty is lost. Liberty guarantees only our outward freedom. If we have liberty but lack free will, then we are only free outwardly but not inwardly. In other words, liberty appears to be empty without free will. To illustrate how this is so, imagine that we live in a world that resembles B. F. Skinner's *Walden Two* (1948). In this scenario, we have the liberty to do as we want, but what we want is the result of careful manipulation that we were subjected to in childhood. The best explanation for why liberty loses its appeal in this example is lack of free will. The actual world that we live in is arguably different. However, determinism is often compared to manipulation by free will skeptics (such as Pereboom, 2014). In denying free will, we consider ourselves to be the same as the residents of *Walden Two*, as far as free will is concerned.

If I am correct that denial of free will erodes motivation because it undermines the perceived meaningfulness of liberty, the remedy for it would be adopting a concept of free

[5] On the flip side, existentialists, such as Sartre, often affirm the value of human existence in affirming free will.

will that highlights agential features enabling liberty to be exercised in ways that are meaningful and cannot be manipulated. While many compatibilists argue that being subjected to deterministic laws is different from being subjected to manipulation by another agent, some compatibilists deny the difference. McKenna (2014), for instance, argues that we should attribute free will to manipulated agents as well, as long as they possess certain agential features, such as reasons-responsiveness. If McKenna is correct, then the kind of free will that we can reasonably hope for is no different from the kind of free will that can be possessed by a manipulated agent. In that case, affirming free will can't save the appeal of liberty after all. In the next section, however, I will argue that the kind of free will that we can reasonably hope for is beyond the kind of free will possessed by a manipulated agent.

It is a frequent complaint that compatibilist notions of free will are shallow[6]. The emotional reaction behind this complaint is disappointment. I think that this emotional reaction should be taken seriously in constructing a concept of free will. As an example, the word "magician" can now be used to refer to people who don't possess supernatural power and therefore don't do real magic; we preserve the concept MAGICIAN even though it contains errors. However, those that can be called magicians are still expected to perform impressive tricks that appear magical. A magician can be called "a magician" only to the extent that they have the appeal of a magician. Similarly, while we may have good reasons to preserve the concept FREE WILL, despite the errors that it contains, what we call "free will" should retain the appeal of free will, if it is to serve as a source of motivation for us.

[6] See Smilansky (2000) for an example.

In response to the complaint that compatibilist notions of free will are shallow, McKenna (2008) argues convincingly that depth is relative and so whether something is shallow depends on what it is compared to. To replace the analogy of depth with ambitiousness, in choosing a concept of free will as a source of motivation, we should seek one that is as ambitious as possible, for what is appealing and motivating is also ambitious. This means that free will should be conceptualized as a capacity that is both valuable and challenging to exercise.

If what I have argued thus far is correct, a concept of free will that we seek as a source of motivation should satisfy two criteria. First, it highlights the agential features that enable liberty to be exercised meaningfully. Second, it is ambitious, i.e., it is valuable and challenging. In the next section, I'll propose a concept of free will that fulfill these two criteria.

3. The free will spectrum

In the free will literature, there has been a long tradition of considering reasons-responsiveness as sufficient for free will[7]. While reasons-responsiveness is no doubt an essential feature of free agents, adopting a concept of free will that amounts to reasons-responsiveness can hardly help improve our motivation. First of all, reasons-responsiveness is not ambitious. On Fischer and Ravizza (1998)'s influential account, for instance, one is considered reasons-responsive if one's counterfactual pattern of actions is "minimally comprehensible" (p.72). Since human irrationality is often comprehensible, one

[7] Examples include McKenna (2013), Sartorio (2016), Fischer and Ravizza (1998), Wolf (1990) and Nelkin (2011).

may act irrationally and still be considered reasons-responsive on Fischer and Ravizza's account. This is an understandable feature of a concept of free will tied to analysis of moral responsibility because we would like to hold people accountable for their irrational actions. However, since every action fulfills the minimal requirement of reasons-responsiveness (or else it is just a bodily movement rather than an action), it is impossible to act without being reasons-responsive. Therefore, regardless of whether we consider ourselves free in the sense of reasons-responsive, when we act, we are always responsive to reasons. Second, while being responsive to reasons is one way in which liberty can be exercised meaningfully, the meaningfulness of liberty is limited for someone who is only reasons-responsive, but not *original*. To clarify what I mean by originality, consider Asch's (1956) famous experiments on social conformity. In the experiment, the participants were presented with three lines of varying lengths and asked to judge which of the three lines is identical to a given line. The answer should have been obvious to all the participants; the test was designed to be easy rather than challenging. Before each participant gave their answer, however, a few experimenters disguised as participants gave obviously erroneous answers. Perhaps not surprisingly, about one third of the participants conformed to the erroneous answer. We may say that those participants were not original.

Are those participants reasons-responsive? They are. Although they answered the experimental question incorrectly, social conformity is comprehensible. In some social contexts, social conformity may even be rational. However, their actions were not guided by their inner source of evidence; rather, they relied on social signals as guidance for actions. When a participant reported that a particular line that obviously looked longer

than another line was identical to that line, the idea that they acted on didn't originate in them; they recycled an idea that they obtained from social observation.

On my account, to act originally is primarily a matter of guiding one's action with one's inner source of evidence. When one acts from what strikes one as evident in one's inner experience, one is acting originally. My account of originality resembles J. S. Mill's, which defines originality as "the capacity of extracting the knowledge of general truth from our own consciousness" (2006, p. 332). As Mill famously argues, an important justification for liberty is to make room for originality. In other words, acting originally is widely recognized as a crucial way in which liberty can be exercised meaningfully. Acting originally doesn't always involve generating novel ideas. We may call originality that doesn't involve novelty *weak originality*. Sometimes, however, acting originally involves generating novel ideas. We may call originality that generates novel ideas *strong originality*.

My account of free will is an extension rather than a replacement of the standard reasons-responsiveness account of free will. On my proposed account, free will occurs on a spectrum. On the least ambitious end of the spectrum, free will is exemplified by actions that are reasons-responsive, but not original. On the most ambitious end of the spectrum, free will is exemplified by actions that are not only reasons-responsive, but also original and novel (strongly original). In the middle of the spectrum, free will is exemplified by actions that are reasons-responsive and original, but not novel (weakly original). In acting, we are always free in the sense of being reasons-responsive; one need not strive to be reasons-responsive. This explains why actions that are only reasons-responsive occupy the least ambitious end of the spectrum. To act originally, on the other hand, requires some effort. Even weak originality is challenging; it requires not only cognitive effort to follow

one's own inner evidence but also (at times) mental strength to resist social conformity. Strong originality is even more challenging, for it requires creative imagination to generate novel ideas. This explains why actions that exhibit strong originality occupy the most ambitious end of the free will spectrum. Weak and strong originality differ only in degree, depending on how novel one's idea is. I think that the degree of novelty of one's idea should be measured by the extent to which the content of one's creative imagination can explain the intelligibility of the novel idea. This allows us to attribute strong originality to someone who improves existing ideas by modifying them. On the other hand, we need not attribute strong originality to those who modify existing ideas in trivial ways that do not improve them. Moreover, what matters to free will is subjective novelty rather than objective novelty. People sometimes independently come up with similar ideas without influencing each other. We should attribute strong originality to those cases as well.

Both weak and strong originality have clear social value; they prevent groupthink. Strong originality in particular results in novel ideas that drive the advancement of humanity. From iPhones to impressionist paintings to tiramisu, prized products of civilization often started as the creation of someone's imagination. As the environment continues to present us with new challenges and sometimes new resources, novel ideas are constantly needed to move our society forward. However, the social value of originality is not the only reason originality should have a place in a concept of free will. I think that a consideration that strongly favors including originality in a concept of free will concerns the first-person phenomenology of free will.

The phenomenology of free will is often considered the primary evidence for it. However, the phenomenology of free will is often absent when we engage in activities that

are reasons-responsive but not original. For instance, paying taxes is a reasons-responsive action. Yet paying taxes isn't normally accompanied by the phenomenology of free will. On the other hand, the phenomenology of free will is often present when we engage in original activities, especially when creative imagination is exercise. It is when we exercise our creative imagination that we experience the full force of human agency.

In addition to contributing to the phenomenology of free will, creative imagination has a special characteristic, which is that its products cannot realistically be manipulated. In principle, an omniscient and omnipotent godlike figure may manipulate how someone's creative imagination proceeds. In reality, however, such manipulation is impossible, for each person's creative imagination is unique and unpredictable. Consider Kant's *Critique of Pure Reason*, which is a prime example of novel ideas. No one could have manipulated Kant to write *Critique of Pure Reason*, because no one else but Kant could have come up with his ideas and expressed them with his style. To be sure, few people in history can create ideas as valuable as Kant. However, this does not entail that my proposed concept of free will is elitist. While some novel ideas, such as Kant's, can make major contributions to humanity, other novel ideas that only contribute to one's personal life are nevertheless valuable. For example, prisoners are well-known for being extremely creative; their creativity can significantly improve their quality of life in a materially-derived environment. Most of us enjoy access to more material resources than prisoners, but when we deeply care about something, we often can't obtain exactly what we want from the market. For instance, someone who deeply cares about coffee often needs to make their own customized coffee that suits their developed taste. As another example, someone who deeply cares about philosophy often needs to write their own philosophical papers that capture their

particular views, which may impact nobody's but their own lives. Sometimes even generating novel but useless ideas can make one more likely to generate novel and useful ideas in the future, simply by exercising one's creative imagination.

Free will is sometimes thought of as requiring metaphysically possible alternatives[8]. While whether such metaphysical possibilities exist is debatable, if someone has such alternatives, first they must conceive those alternative possibilities. Someone who can come up with novel ideas for actions from their creative imagination is more likely to conceive of alternative possibilities to their actions than someone who can't. Thus, someone who is strongly original is more likely to have metaphysically possible alternativesto their actions. Even if free will does not require metaphysical alternative possibilities, someone who can't conceive of alternative possibilities to their actions in a given situation is likely to feel trapped and unfree.

My account of free will can also be seen as providing a ultimacy condition on free will, which will be discussed in the next section.

4. Originality and ultimacy

On a popular view on free will, free will has a ultimacy condition[9]—one is free when one's action ultimately originates in oneself. My account can be thought of as providing such a ultimacy condition. On my proposal, one's action ultimately originates in oneself to the extent that one acts with strong originality, that is, when one acts on an original and novel idea.

[8] Notable examples include Kant (1788) and Ayer (1972).
[9] Notable examples include Galen Strawson (1994) and Smilansky (2003)

My account of ultimacy captures our everyday intuition. When we ask, ultimately, who came up with the original idea of pizza as we know it? The answer is the Italians but not the Americans, even though the Americans who make pizzas do so following their inner source of evidence. This shows that in everyday conversations, ultimate origination requires not just weak originality, but also strong originality.

In defending the compatibility of determinism and ultimacy, McKenna (2008) similarly appeals to the ordinary meaning of ultimate origination. McKenna asks us to consider the following case:

> Now imagine that I have a group of friends with me on one of my insanely long, mountainous and especially taxing bicycle rides. One of the crew cries out, "Whose hair-brained idea was that?" It might have been various members of the group discussed the matter with me after I proposed it, entertained alternative routes, and so on. But if I was the one who hatched it, I can, so far as I can tell, fess up and say that it ultimately originated with me when I was studying a map prior. I am the ultimate source of that plan and the fine mess that I have gotten the entire group into. Could I really reply, "Well, determinism being true and all, though it was my idea, ultimately it really did not originate with me—it started with the big bang?" (p.198)

According to McKenna's analysis, the hare-brained idea in the story ultimately originated in the person who "hatched" it, that is, the person whose imagination created this novel idea. This is consistent with my account of ultimacy as strong originality. Interestingly, although after reading the story above, one might expect McKenna to equate ultimacy with originality, he proposes two accounts of ultimacy that are very different from mine. While I

agree with McKenna that ultimacy can be compatible with determinism, in the following, I will argue that my account of ultimacy is preferable to McKenna's.

McKenna's first account of ultimacy is tied to Fischer's expressive theory, according to which the value of our agency is understood in terms of a narrative. On this theory, the value of one's action is derived from its contribution to one's life narrative. McKenna argues that while not all one's actions contribute to one's life narrative, those that "arise as distinct contributions" (p. 200) to one's life narrative can be understood as ultimately originating in oneself.

My criticism of McKenna's first account of ultimacy consists of two points. First, adopting a concept of free will that requires seeing our lives as narratives has little action-guiding value. We are not protagonists of Hollywood movies, and whether one's life makes a good narrative is quite independent of whether one's life is good. While we can specify the evaluative criteria for life as a narrative such that a good narrative of life can only come from a good life, seeing one's life as a narrative doesn't seem to be a particularly useful cognitive tool to me[10]. Second, actions that arise as distinct contributions to one's life narrative may not reflect any deep features of one's agency. For instance, John Nash was diagnosed with schizophrenia, and the actions that arose from his schizophrenic symptoms significantly contributed to his life narrative. In fact, when Nash's life is portrayed in the movie *A Beautiful Mind*, a significant part of the narrative is about his schizophrenia. Did Nash's schizophrenic symptoms ultimately originate in Nash? In a superficial sense, they did. However, they did not originate in Nash in a deeper sense, as the Nash equilibrium did, for instance. Schizophrenia aside, mental illnesses in general often dramatically contribute

[10] For a similar viw, see Galen Strawson (2004); for an opposite view, see Schechtman (2014).

to one's life narrative against one's wishes. The relevant actions here often reflect only superficial features of one's agency, and lack the sort of depth that we can reasonably hope for in human freedom.

McKenna's second account of ultimacy is tied to the practice of attributing moral responsibility. In a moral community, members attribute meanings to each other's actions and express their expectations to each other, sometimes by expressing reactive attitudes. "If one's actions can be fitted into a fabric of interpersonal expectations in our moral interactions… then her actions can be seen as ultimately originating in her contribution to the fabric" (p. 202). What is it like to be fitted into a fabric of interpersonal expectations? A story from Aesop's Fables, titled "The man, the boy, and the donkey" provides some insight into the phenomenon. The story goes like this. A man and a boy go to the market with a donkey. They meet a stranger on the way, who criticizes them for not riding the donkey. So the boy rides the donkey. They then meet another stranger who criticizes the boy for not letting his father ride the donkey. The boy then gets off the donkey and lets his father ride it. Then another stranger criticizes the father for not letting his son ride the donkey. So both the man and his son get on the donkey. They then meet another stranger who criticizes them for abusing the donkey. The story goes on like this and eventually, the donkey dies. While we can still attribute free will to the man and the boy, their effort to fit into a fabric of interpersonal expectation does not seem to be a source of it. We can again consider the participants in Asch's experiment. When they conformed to the erroneous answer previously presented, no doubt they were aware of what might be expected from them. It was due to this assumption of what might be expected from them socially that they did not act on their original ideas. In other words, the sort of social awareness and

calculation that underlie McKenna's second account of ultimacy is sometimes an obstacle to originality. Moreover, this sort of social awareness and calculation does not always result in pro-social actions. To the extent that we are responsive to praise and blame, we care about only praise and blame from those who we consider to be a part of our moral community. In practice, this responsiveness to praise and blame from in-group members sometimes results in aggression towards out-group members. Think of Nazi Germany, for example; ordinary Germans who joined the Nazi Party did so not because they were not responsive to praise and blame, but because they were responsive to praise and blame from the Nazis.

Responsiveness to praise and blame is an even more obvious obstacle to strong originality. If I act on received ideas, I know what praise or blame I can reasonably expect. However, if I act on a genuinely novel idea, I do not know how other people may react to it; I may be ridiculed. Numerous studies conducted by different researchers show that people tend to be less creative when they expect that their creative products will be evaluated (e.g., Hennessey, 2001; Shalley, 1995; Amabile, 1979; Amabile, Goldfarb & Brackfleld, 2009). This may partly explain why, even in academic philosophy, many publications are slight variations on received ideas rather than expressions of genuinely novel ideas. On the other hand, generation of novel ideas seems to require a degree of indifference to social reactions. This explains why many artists, such as movie directors, make a deliberate decision to refrain from reading reviews of their works. A quote from Guston, cited by Dennett (2001) and Lehrer (2012), highlights the importance of blocking the sort of social awareness and calculation that McKenna takes to be central to free will in creative activities: "When I first come into the studio to work, there is this noisy crowd which

follows me there; it includes all of the important painters in history, all of my contemporaries, all the art critics, etc. As I become involved in the work, one by one, they all leave. If I'm lucky, every one of them will disappear. If I'm really lucky, I will too."

Given the conflicts between originality and conformity to social expectation, which condition is more central to ultimacy? I believe that the answer is the former. When I act on a novel idea that originated in my imagination, it is my imagination, but not others', that provides the primary explanation for my action. On the other hand, when I act on a received idea, perhaps out of awareness of what other people expect of me, I am but a mediated source of my action, that ultimately originated in someone else's imagination. It is for this reason that my account captures the ultimacy condition better than McKenna's.

5. Originality and desert

As I mentioned earlier, I think that moral responsibility is overemphasized in contemporary discussions of free will. However, I think that those who theorize about free will in terms of moral responsibility are apt to connect free will with desert. We want to deserve what we have in life, and free will appears to be a necessary condition for it. In this section, I will discuss the connection between my account of free will and desert.

While we may care about deserving both the praise and blame that we receive, we really care more about receiving deserved praise then deserved blame. But even deserved praise from other people is of little value. As Hume puts it, "a gloomy, hare-brained enthusiast, after his death, may have a place in the calendar; but will scarcely ever be admitted, when alive, into intimacy and society, except by those who are as delirious and dismal as himself" (2006, p.258) What we should strive for, then, is not to receive deserved

praise from other people, but to be included in their community. If we are ambitious, we should strive to be included in their conception of utopia. If we are even more ambitious, we should strive to have an irreplaceable position in their concept of utopia.

My idea of utopia resembles Nozick's (1968). On his account, the meta-utopia consists of many different utopias that appeal to people with different interests. Each utopia has a filtering process; a person can only stay in a utopia if other members of the utopia agree to it. By design, no one fits in all utopias because communities residing in different utopias value different kinds of members. Thus, free will on my account is not a sufficient condition for being included in all utopias. However, free will on my account is a necessary condition for having an irreplaceable position in some utopias. How so?

Whether membership in a utopia is granted depends on what value one can contribute to the utopia. In the past, physical labor was essential for the thriving of a community; people could contribute to a community significantly by acting on other people's ideas, such as by building the pyramids. Today, artificial intelligence is more advanced than ever; soon physical labor from people will no longer be valuable. Someone who can significantly contribute to a utopia, then, is someone who can come up with novel ideas, that is, someone with strong originality. Original ideas are usually unique. Thus, when someone comes up with original ideas, their contribution is usually unique and thus irreplaceable. On the other hand, someone who is responsive to praise and blame but is not original is replaceable.

Adopting my account of free will not only enables one to deserve an irreplaceable position in other people's conception of utopia, it also enables one to deserve something even more valuable, namely, liberty. Most of us do not deserve liberty in a backward-

looking sense, since we did nothing to earn it. However, by adopting my concept of free will and exercising liberty in meaningful ways, we can deserve liberty in a forward-looking sense[11]. When compared to those who do not exercise liberty meaningfully, those who do deserve liberty maximally.

6. Conclusion

In this chapter, I have proposed an account of free will that extends the traditional reasons-responsiveness accounts. On my account, free will occurs on a spectrum. On the least ambitious end of the spectrum, free will is exemplified by actions that are reasons-responsive, but not original. On the most ambitious end of the spectrum, free will is exemplified by actions that are not only reasons-responsive, but also original and novel. In the middle of the spectrum, free will is exemplified by actions that are reasons-responsive and original, but not novel. Free will is exemplified by different types of actions because they may be required by different circumstances. Under some circumstances, reasons-responsiveness alone suffices for free will. Under other circumstances, an ambitious free agent would strive for not only reasons-responsiveness, but also originality.

While my discussion is divorced from analysis of moral responsibility, adopting my account of free will enables one to deserve an irreplaceable position in others' conception of utopia as well as liberty.

[11] See Schmidtz (2002) for an example of a forward-looking account of desert.

Chapter Two

Luck in Rational Freedom:

An Argument for the Compatibility between Free Will and Agent-Internal Luck

> *Luck and control are at the heart of important unfinished business for all believers in free, morally responsible agency.*
> —Alfred Mele, Free Will and Luck, 2006

1. Introduction

In this chapter, I will argue for the compatibility between free will and agent-internal luck, based on the common compatibilist assumption that free will-relevant control is rational control. Some compatibilists—notably Mele (2006) and McKenna (draft)—think that while agent-internal luck is compatible with free will, it is nonetheless an undesirable feature that a free, rational agent would strive to eliminate. Call this view restrictive compatibilism about agent-internal luck. Against this view, I will argue that optimal rational control is compatible with agent-internal luck. Moreover, I will argue that more practically rational agents tend to exhibit more agent-internal luck in their decision-making processes.

Luck is often seen as an enemy to free will. Some philosophers, such as Levy (2011), deny the very possibility of free will on the grounds that all of our actions are ultimately just a matter of luck. Even free will realists often reject an account of free will because it subjects a supposedly free decision to some unacceptable form of luck (see, for instance, O'Connor, 2000; Mele, 2006). The most frequently rejected form of luck is what I will call agent-internal luck, which results from psychological indeterminism that occurs in one's decision-making process. To borrow an example from Kane (1996), consider a

businesswoman who is torn between helping someone in need and hurrying off to an important meeting. Suppose the woman decides to help the person in need at t. Suppose also that holding fixed everything that happened before t, the woman may instead decide to hurry off to her meeting at t. In this case, nothing about the agential features of the woman or the circumstance that she is in prior to t can explain why she decides to help the person in need instead of hurrying off to her meeting at t. Her decision to act one way rather than another at t, it seems, is just a matter of luck.

Since the existence of agent-internal luck is more obvious if one's decision-making process is physically indeterministic, agent-internal luck has long been considered a problem exclusively for libertarianism. Consequently, Hume notoriously argues that free will requires the truth of determinism[12]. But if agent-internal luck is a problem for libertarians, it is a problem for contemporary compatibilists as well[13], and for two reasons. First of all, unlike Hume, contemporary compatibilists typically do not presuppose the truth of determinism, and thus their accounts of free will are not immune to the possibility of indeterminism. Indeed, for all that we know, physical indeterminism is a genuine possibility. Second, although less obviously, one's decision-making process can be psychologically indeterministic even if it is physically deterministic[14]. This is so because an identical psychological state (which occurs at the macroscopic level) can be realized by multiple distinct neurological states (which occur at the microscopic level), which can subsequently lead to multiple distinct psychological states (decisions). Thus, inputs to one's decision-making process that are neurologically distinct but psychologically

[12] Hume (2011)
[13] See Vargas (2012), de Calleja (2014), McKenna (draft) for a similar point.
[14] List (2014) makes a similar point. Ismael's (2016) treatment of determinism can also illustrate the possibility of psychological indeterminism in the context of physical determinism.

indistinguishable can deterministically cause one to make different decisions. To illustrate this idea using Kane's example again, imagine that the macroscopic, psychological state of the businesswoman just before t (at t') can be realized by either the neurological state N1 or a slightly different neurological state N2. In one possible world, the woman's psychological state at t' is realized by N1, which deterministically causes the woman to decide to help someone in need at t. In another possible world, her identical psychological state at t' is realized by N2, which deterministically causes her to decide to rush off to the meeting at t. In this case, physical determinism does not preclude psychological indeterminism. From the perspective of the businesswoman, the selfsame psychological state at t' may cause her to make either decision at t. Explaining the compatibility between free will and agent-internal luck, therefore, is a challenge that confronts everyone who believes in the possibility of free will. My objective in this paper is to take up this challenge.

To understand the nature of the challenge, the most important question to ask is, what is so problematic about agent-internal luck anyway? Answers from the literature differ in detail, but all point toward some alleged conflict between agent-internal luck and agential control.[15] According to those who regard agent-internal luck as antagonistic to free will, if psychological indeterminism occurs at a person's decision-making process, then the person lacks some sort of control (call it "unique control") over the outcome of the decision-making process. To put it slightly differently, a person exercises unique control over a decision when the decision is uniquely favored by the person's agential features

[15] Levy (2011), for instance, expresses the concern as follows: "If it is a matter of luck whether the agent performs one action rather than another, then the agent seems to lack the kind of control over his actions required for genuine freedom." (p.45) Similarly, Mele (2006) writes, "[W]hen luck (good or bad) is problematic, that is because it seems significantly to impede agents' control over themselves or to highlight important gaps or shortcomings in such control." (p.7)

prior to the decision-making process that issues the decision. If a person lacks unique control over a decision, the argument goes, then the person doesn't make the decision freely. But even if free will involves some necessary control conditions, it is unclear why unique control is among those conditions.

A brief analysis of different attitudes towards the relation between free will and control will prove illuminating here. The most metaphysically demanding control condition is what Fischer (2006) calls "total control"[16], which is characteristically endorsed by free will skeptics. Compatibilists and libertarians alike deny that total control is a necessary condition for free will. On the other hand, libertarians tend to think that free will requires—among other things—what Kane (2005) calls "plural voluntary control"[17], which allows a person to make multiple rational decisions in response to one and the same practical problem, even holding fixed the person's agential features prior to the decision-making process that issues the decision. Notice that by definition, plural voluntary control is only possible when unique control is absent. Most compatibilists—typically those who are convinced by Frankfurt's argument against the Principle of Alternate Possibilities—deny that free will requires either total control or plural voluntary control.

Now if free will requires total control, then free will may require unique control. However, if free will requires total control, then one never decides freely anyway, regardless of whether unique control is present or absent, since total control is impossible

---

[16] Fischer defines total control this way: "An agent has total control over *X* only if for any factor *f* which is a causal contributor to *X* and which is such that if *f* were not to occur, then *X* would not occur, the agent has control over *f*." (p.116) Total control resembles what Levy (2011) regards as "relevant control" (p.5).

[17] According to Kane, people exercise plural voluntary control when, facing different decision options, they "are able to bring about whichever of the options they will, when they will to do so, for the reasons they will to do so, on purpose, rather than accidentally or by mistake, without being coerced or compelled in doing so or in willing to do so, or otherwise controlled in doing or in willing to do so by any other agents or mechanisms." (p.138)

to have. On the other hand, if free will requires unique control rather than plural voluntary control, this can be so only if unique control is somehow more important than plural voluntary control for one's free will, since unique control and plural voluntary control cannot be exercised simultaneously. How can unique control be more important than plural voluntary control for one's free will? Consider the compatibilists who argue against the necessity of plural voluntary control for free will. Their arguments tend to emphasize the rational aspect of agency. Their accounts of free will tend to hold that one decides freely to the extent that one's decision results from certain agential features that mark one as a rational agent.[18] In other words, arguments against the necessity of plural voluntary control are typically based on the following assumption: Rational control is a sufficient control condition for free will. If this assumption is correct, then free will requires unique control only to the extent that rational control requires unique control.

My target audience consists of those who share the assumption that rational control is a sufficient control condition for free will, i.e., those who deny that free will requires either total control or plural voluntary control. As I will argue, free will does not require unique control—that is, free will is compatible with agent-internal luck—because rational control does not require unique control. Moreover, I will argue that the general existence of psychological indeterminism is an essential feature of rational agency, even though it may be absent in any given instance of decision-making. Some philosophers, notably Mele (2006) and McKenna (draft), think that free will is compatible with agent-internal luck, but their view has the implication that a rational agent would nevertheless strive to restrict the

[18] Examples include McKenna (2013), Sartorio (2016), Fischer and Ravizza (1998), Wolf (1990) and Nelkin (2011).

possibility of agent-internal luck. Call this view "restrictive compatibilism about agent-internal luck", which I will criticize (in §2) before proceeding to my main argument (in §3).

Let me first clarify some ambiguous terminology that appears frequently in this paper. Following Mele (2006), I use "deciding" to refer to the mental act of forming an intention to perform (or refrain from) a particular action. On the other hand, I use "decision-making process" to refer to the entire psychological process that underlies one's attempt to make a decision in response to a practical problem, which may involve deliberating and deciding. Consequently, agent-internal luck, as defined in the beginning of this paper, may occur both at the deliberating stage and at the deciding stage. Agent-internal luck occurs at the *deliberating stage* in the following kind of situation: When holding fixed the external conditions, a person's (macroscopic) agential features prior to deliberation do not provide a contrastive explanation for the deliberative outcome. On the other hand, agent-internal luck occurs at the *deciding stage* in the following kind of situation: When holding fixed the external conditions, a person's (macroscopic) agential features prior to a decision's being made do not provide a contrastive explanation for the decision. Both kinds of agent-internal luck will be discussed in this chapter. Finally, one exercises rational control over a decision when the underlying decision-making process is a rational response to the practical problem that prompts the beginning of the decision-making process.

2. Restrictive compatibilism about agent-internal luck

2.1. Mele's version

Mele (2006) offers two arguments for the compatibility between free will and agent-internal luck, which are correspondingly based on his two libertarian accounts of free will—modest libertarianism and daring soft libertarianism. According to modest libertarianism, a decision may remain free when a consideration that is "indirectly relevant to a deliberator's present practical question comes to mind during deliberation but was not deterministically caused to do so" (p.12), even if different indeterministically caused considerations may lead one to arrive at different deliberative outcomes. According to daring soft libertarianism, on the other hand, a decision may remain free when it is indeterministically caused by one's deliberative outcome. In other words, modest libertarianism allows the compatibility between free will and agent-internal luck that occurs at the deliberating stage; while daring soft libertarianism allows the compatibility between free will and agent-internal luck that occurs at the deciding stage.

I will not repeat Mele's arguments for either of his libertarian accounts, since they are largely irrelevant to the present discussion. What I want to highlight here is Mele's general attitude towards agent-internal luck. In repeatedly using the phrases "good luck" and "bad luck" to describe different possible outcomes of an instance of agent-internal luck, Mele gives his readers an impression that agent-internal luck always subjects one to the possibility of making rationally suboptimal decisions. If so, then agent-internal luck remains problematic for one's rational control despite its compatibility with free will, for a rational person would strive to restrict the future possibility of making a decision that amounts to bad luck. Indeed, Mele explicitly recommends that one reduce the likelihood that agent-internal luck occurs at the deciding stage, in the following passage:

> … DSLs (daring soft libertarians) seek, as agents, to eliminate luck in an important sphere. To return to the image of the roulette wheel, they seek to replace an indeterministic wheel connecting considered judgment and action with a deterministic one. (p.117)

Similarly, Mele suggests some strategies for modest libertarian agents to "override" (p.123) agent-internal luck that occurs at the deliberating stage, such as "by engaging in further deliberation" (p.123). By spending more time on deliberating, according to Mele, one can reduce agent-internal luck that results from not taking a relevant consideration into account during deliberating. As I will argue later (§3.2), although taking more relevant considerations into account during deliberating can reduce the likelihood that agent-internal luck occurs at the deliberating stage (as Mele suggests), doing so can increase the likelihood that agent-internal luck occurs later at the deciding stage (which Mele considers undesirable). In other words, Mele's advice cannot help one reduce the overall prevalence of agent-internal luck. As I will also argue, however, the persistence of agent-internal luck should not be regarded as problematic for one's rational control. While Mele's writing projects an image of agent-internal luck that resembles influenza—an inevitable part of life that one should always guard oneself against—this is a distorted image of agent-internal luck.

2.2. McKenna's version

McKenna (draft) focuses specifically on agent-internal luck that occurs at the deciding stage. Unlike Mele, McKenna also considers the possibility of deterministic agent-internal

luck, which, according to McKenna, occurs when a decision is the result of conflicting motivations with roughly equal strengths. Like Mele, however, McKenna considers the occurrence of agent-internal luck to be problematic for one's free will. According to McKenna, a decision that is a direct outcome of agent-internal luck remains free only if it is an indirect outcome of one's prior decision that did not itself directly result from agent-internal luck. Two remarks are in order here.

First, agent-internal luck is much more pervasive than McKenna appears to assume. McKenna only considers the possibility of agent-internal luck that results from conflicting ends, but there are other factors that can also lead to existence of agent-internal luck in a decision-making process. For instance, agent-internal luck may occur when one faces a choice between conflicting means to one's end; our everyday life is full of "Buridan's Ass"-type cases where one can achieve the same end by employing different means that are equally good or incommensurable. In Kane's original example, the businesswoman faces the dilemma of whether to realize her general end "to help someone in need" by offering help to someone she encounters on her way to an important meeting. As Singer (2016) famously points out, we face a similar dilemma every day if only we care to consider those people in need who are not physically proximal to us. In everyday life, one's general end to help someone in need can be realized by numerous means—there are numerous charity organizations that one can make a donation to or volunteer for. One may even choose to establish one's own charity organization. In other words, not only does one face a dilemma similar to that of the businesswoman in Kane's example on an everyday basis, one is also unlikely to uniquely favor any particular action, even if one prioritizes one's general end to help someone in need over any other ends. Agent-internal luck can also arise when a

decision is based on insufficient information about relevant empirical facts, especially facts about the future. Our decisions are future-oriented, in the sense that they aim to bring about certain states of affairs in the future. The further into the future a decision concerns, the more relevant but inaccessible empirical facts there are[19], and thus, the more uncertainty is involved in the decision-making process. An important source of uncertainty concerns the nature of one's future self; L. A. Paul (2014) uses the example of deciding whether or not to have children to highlight the unpredictability of how one's agential structure may change after what one makes what she calls a "transformative decision." In the face of uncertainty about the relative merits of different options, a rational person is less likely to uniquely favor any particular decisions, and more likely to judge different options as either equally good or incommensurable. In the next section, I will explain in more detail as to why I think that a rational person's decision-making process often involves agent-internal luck. If I am correct about the pervasiveness of agent-internal luck, McKenna's view has the theoretical consequence that one's free will is severely restricted. This is a risky (and unnecessarily risky) strategy for a free will realist.

Second, McKenna's view is based on an implicit assumption—which is shared by Mele and is common in the literature—that one is always less than rationally optimal when one may act in random ways even after deliberation[20] (that is, when agent-internal luck occurs at the deciding stage). This mistaken assumption, I suspect, is sustained by philosophers' tendency to focus on cases where one may act randomly even after deliberation and is less than rationally optimal. There are two kinds of situations where agent-internal luck may occur at the deciding stage: when one's decision deviates from

[19] See Lenman (2000) for a similar point.
[20] See Sartorio (2021) for an exception.

what is uniquely favored by one's deliberation, and when one's deliberation does not uniquely favor a particular decision. Examples of the former in the literature tend to be cases of akrasia, where one is uncontroversially less than rationally optimal. On the other hand, examples of the latter in the literature are typically cases where one is torn between a moral decision and an immoral decision. In such cases, someone whose deliberation does not uniquely favor the "correct" decision—namely, the moral one—may strike us as less than rationally optimal. However, in everyday life, even rationally optimal deliberation often does not uniquely favor a particular decision. In cases where rational deliberation does not uniquely favor a particular decision, agent-internal luck should occur at the deciding stage if one is rational. I will elaborate more on this point in the next section.

3. My solution to the agent-internal luck problem

In this section, I will argue that optimal rational control is compatible with both agent-internal luck that occurs at the deliberating stage and agent-internal luck that occurs at the deciding stage. As I will explain further, there often isn't a uniquely correct solution to a practical problem, especially in a supportive environment. The longer one deliberates rationally in response to a particular practical problem, the more likely one is to end up with different options that strike one as incommensurable, which results in agent-internal luck that occurs at the deciding stage. Recognizing this, one may rationally shorten the amount of time one engages in deliberation by taking only into account fewer and randomly selected options. Sometimes, the rational response to a practical problem is to satisfice—that is, to stop deliberating as soon as one finds a good enough option. In such

cases, agent-internal luck that occurs at the deliberating stage is compatible with optimal rational control.

A possible response to my argument is that one's rational control is always suboptimal when agent-internal luck occurs. How this response may develop partly depends on how optimal rational control is conceptualized. If optimal rational control is beyond what an actual person can achieve—for instance, if it requires total control—then one's rational control is suboptimal when agent-internal luck is present, but also when it is absent.

On the other hand, if optimal rational control is set by the limits of what an actual person can achieve, then it is incompatible with agent-internal luck only if at least one of the following statements is true: (1) a particular decision-making process is always a less rational response to a given practical problem when agent-internal luck is present than when it isn't; (2) someone who is in general more prone to agent-internal luck is less rational than someone who isn't. These two statements correspondingly concern the nature of rationality both at the decision-making level and at the agential level, each of which will be discussed below.

3.1. Regarding rational decision-making

Let us first examine (1), the claim that a particular decision-making process is always a less rational response to a given practical problem when agent-internal luck is present than when it isn't. (1) is true only if, despite what I argued above, (a) there is always a uniquely rational deliberative outcome in response to a given practical problem (such that the occurrence of agent-internal luck at the deliberating stage always indicates flaws in the

way one deliberates), and (b) that the outcome of rational deliberation always uniquely favors a particular option (such that agent-internal luck only occurs at the deciding stage when one decides against one's deliberative outcome).

Let us first examine (b), the assumption that the outcome of rational deliberation always uniquely favors a particular option. There are many philosophical accounts of practical reasoning. On most of them, rational deliberation need not uniquely favor a particular option. Below, I will make my case by considering two of the most popular accounts—expected utility theory and Kantian reasoning.

Expected utility theory is implicitly endorsed by Levy (2011), who has developed the most comprehensive luck-based argument for free will skepticism:

> in the process of deliberation the agent discovers the weights that her reasons have for her, where this is a weight they had for her all long but which she was unable to perceive clearly. Plainly, this kind of discovery is a typical upshot of careful deliberation. (p.68)

On Levy's account of deliberation, each of one's reasons has an objective and stable weight, which can be represented by a numerical value. The task of rational deliberation is to discover which option under consideration is supported by one's weightiest reason(s). This line of reasoning can be extrapolated to support (b)—If each option under consideration can be assigned a numerical value, then rational deliberation should uniquely favor the option associated with the highest value. This resembles expected utility theory, which states that a rational person always chooses a course of action expected to result in a state of affairs with the highest utility.

In response, notice that even if expected utility theory is correct, rational deliberation need not uniquely favor a particular option. Through rational deliberation, one may discover that different options have the same expected utility; that is, there exists a tie. If there is a tie, then the outcome of one's rational deliberation is indifferent between different options. Moreover, expected utility theory is silent on what choice one should make when one's preferences are not ranked, such as when one encounters incommensurable options, which cannot be measured against each other on a common scale.

Consider under what conditions there exists a single scale on which all the relevant goodness embodied in each option can be measured against any other, from one's perspective. This can only be the case when one considers one's different available options to be good in only one way, perhaps because the environment severely limits what options one has, or because one fails to appreciate the variety of ways in which things may be good. In our contemporary world, the first factor is often, though not always, irrelevant; we have numerous options almost in every aspect of life. A good example is career options. Compared to our ancestors, we are overwhelmed with the multitude of career options that we can choose from. Consider the case where one needs to decide which career to pursue. For some people, the choice is relatively easy—they consider relevant only the expected financial return from a career. For those people, pursuit of a career in academic philosophy, for instance, is unthinkable. Different career options are commensurable for them, but only because they neglect a variety of ways in which a career option may be good. In other words, in a supportive environment, commensurability results from neglect[21]. On the other

[21] Millgram (2005) makes a similar point.

hand, if one appreciates a wide variety of ways in which a career option may be good, one is more likely to end up considering different options incommensurable—for instance, one option may help develop one's ability of abstract thinking, another scientific reasoning, another aesthetic sensibility, yet another interpersonal skills, etc. The list can go on forever. This is true not only for big decisions such as deciding what career options to pursue, but also for decisions as small as deciding what ice cream to eat. Making ice cream at home allows me to control what ingredients go in it, but it is time consuming. Buying ice cream from a grocery store is convenient, but the ice cream often contains questionable ingredients. McDonald's ice cream provides nostalgic value, but its taste lacks complexity. Vegan ice cream is animal-friendly, but it lacks the creamy texture of diary ice cream. Gelato contains fresh ingredients, but I would be confronted with the problem of incommensurability once again when I am asked to choose from the wide variety of unconventional flavors that a gelato shop tends to offer. Turkish ice cream has a unique chewy texture that I like, but I would need to travel outside Tucson, where I live, to get some. Again, the list can go on indefinitely. The longer and more carefully one deliberates about the matter, the more relevant but incommensurable kinds of goodness one is likely to discover during deliberation, which prevents one from uniquely favoring any particular option at the end of deliberation. Agent-internal luck thus occurs at the deciding stage.

Let us now consider why Kantian reasoning need not uniquely favor any particular action either. Kantian reasoning only allows one to exclude certain actions, that is, those that violate the categorical imperative. But many actions are consistent with the categorical imperative. Consider Kane's example of the businesswoman again. Kantian reasoning informs one that helping someone in need is an imperfect duty, but developing one's

talents, which can be achieved by going to an important meeting, is also an imperfect duty. When different imperfect duties are in conflict, Kantian reasoning is silent on which imperfect duty one should choose to fulfill. More obviously, Kantian reasoning is silent on what decisions to make on non-moral matters—such as what career to pursue, or what ice cream to eat after dinner. In other words, according to both expected utility theory and Kantian reasoning, rational deliberation need not uniquely favor any particular options. This shows that on most widely accepted views on practical reasoning, agent-internal luck may occur at the deciding stage after rational deliberation. The burden of proof is thus on my opponents.

I will now turn to (a), the assumption that there is always a uniquely rational deliberative outcome in response to a given practical problem. I argued above that for decisions big and small, in a supportive environment, the longer and more carefully one deliberates, the more likely it is that one's deliberation concludes with different options that one deems to be incommensurable or ties. As Buss (2020) observes, when we describe any action in enough detail, there exists an alternative action that we find equally preferable. For example, when one walks up the stairs, one often finds it equally preferable to plant one's left foot in one way and in a slightly different way. Prolonged deliberation therefore cannot eliminate incommensurability or ties. What a rational person should do, then, is not engage in deliberation for as long as possible. Rather, what a rational person should do is to satisfice—that is, to choose an option that one deems as good enough[22]. Even in situations where prolonged deliberation can lead one to find an option that one considers the best, satisficing is a rational strategy to control costs of deliberation.

[22] Proponents of satisficing include Slote (2001) and Simon(1997).

Sometimes, satisficing is a rational strategy due to time constraints. When one does not have the time to weigh the cost and benefits of each option, it is better to satisfice than not make a decision at all. Even when there is no time constraint, further deliberation may not improve the quality of one's decision. Baron, Beattie and Hershey (1988) found that people exhibit what they call "information bias"; that is, people tend to believe that the more information they have when making a decision, the better their decision will be, even when the extra information is irrelevant and distracting. Similarly, Sivanathan and Kakkar (2017) found that when people consider the serious side effects of a drug along with other side effects of the drug, their judgements of the overall severity of the drugs are "diluted" and less accurate, compared to considering only the serious side effects of the drugs.

If satisficing is a rational strategy, there need not be a unique deliberative outcome in response to a given practical problem, so long as one's deliberation concludes with an option that one deems to be good enough. After all, there are often multiple options that are good enough for solving a particular practical problem. Levy (2011) writes that "insofar as we are rational agents, a non-random selection of considerations is likely to cross our minds, with the selection constrained by our values and beliefs." (p.92) But if satisficing is a rational strategy, it is rationally permissible to take into account a random selection of considerations during deliberation, so long as they allow one to find a good enough option. When deciding what career to pursue, for instance, I can take into account only a random selection of career options during deliberation, so long as the deliberation concludes with a career option that I find good enough. In game theory, a mixed strategy is sometimes used in attempt to reach an equilibrium with other actors. Since a mixed strategy requires randomization of options, randomization of one's deliberative outcome is desirable when

one uses a mixed strategy[23]. In other words, (a) is false; agent-internal luck may occur at the deliberating stage even if one is fully rational.

Is it possible that for any given practical problem, there really is always an optimal solution, but even a fully rational person may not be able to find it through deliberation? Two remarks are in order here. First, suppose one deliberates in a fully rational manner and yet one's deliberation does not conclude with the optimal solution to one's practical problem; that deliberative outcome should not as a result count as rationally suboptimal. Second, the normative value of a decision need not remain stable. Sometimes, the normative value of one's decision increases as a result of one's endorsing it. Nietzsche provocatively suggests that one can and should validate each of one's decisions by endorsing it[24]. More modestly, Schmidtz (1994) points out that one's choice to pursue an end often changes one's attitudes and subsequent actions so as to make the end more valuable to one. In other words, one can increase the normative value of one's decision through future actions. For instance, if, after deciding to become a lawyer instead of a painter, I devote most of my energy to painting rather than studying law, my reason to become a lawyer would be weaker than if I fully commit to working towards a career in law afterwards. This point is related to the characteristics of a rational person, which will be the focus of the next section.

3.2. Regarding rational agency

[23] See Binmore (2007).
[24] Nietzsche (1947) proposes that we should embrace the eternal recurrence of each event in our lives.

Let us now examine (2), the claim that someone who is in general more prone to agent-internal luck is less rational than someone who isn't. This view is implicit in Mele's suggestion that one should strive to minimize agent-internal luck that occurs both at the deliberating stage and at the deciding stage. We can see from the discussion above why this view is misguided. The longer one engages in rational deliberation, the more likely one is to end up with different options that strikes one as incommensurable or ties, in which case agent-internal luck occurs at the deciding stage. On the other hand, if one satisfices and stops deliberating as soon as one finds a good enough option, agent-internal luck is likely to occur at the deliberating stage. In fact, as I will argue below, a more practically rational person tends to exhibit more agent-internal luck.

Consider what a practically rational person is like. Minimally, it is someone who excels at solving practical problems. Here, I will only highlight two agential features that facilitate practical success—possession of (a) numerous means, and (b) numerous ends.

I will begin with (a), which is rather straightforward. One kind of practical problem that people constantly face concerns finding means to attain some particular end. Since what means one can employ to attain a particular end is restricted by the external environment, to be able to meet various unpredictable external demands, a practically successful person is someone with plentiful intellectual and material means.[25] When the environment is supportive, a resourceful person needs to choose from an extended menu of means. The more means one possesses, the more likely one possesses means that are incommensurable or equally good for a given situation, in which case agent-internal luck tends to occur. Consider payment methods, for instance. A practically rational person in the

[25] Millgram (2005) makes a similar point.

contemporary world likely possesses numerous payment methods: cash (perhaps of different currencies), credit cards, a cheque book or two, a PayPal account, etc. When the person purchases something from a merchant that accepts all of these payment methods, he is likely to pick an option randomly. What kind of external environment one faces often depends on how successful one is at solving practical problems as well. Consequently, a practically successful person is also less likely to get into situations where he is forced to choose from a limited number of means to attain his ends.

Let us now move on to (b). Although it is sometimes assumed in the literature that solving practical problems only involves coming up with means to attain some pre-existing ends, this assumption does not stand up to scrutiny. [26] In the past, most people were preoccupied with the largely shared urgent end of survival along with other ends pre-assigned by their society. Nowadays, a common practical problem that many people face concerns what ends to attain with their available means, such as time and financial resources. A practically rational person, who appreciates the wide variety of goodness that life can offer, tends to possess numerous ends and be able to acquire new ones quickly. Consider someone on a long flight. Someone with limited interests is prone to wasted few precious hours. While reading is often a good option, this is so only if one is interested in reading and is blessed with the good fortune of not having a crying child on board. Similarly, the limited entertainment options that are usually available on modern planes are only real options to those with sufficiently broad interests in movies or music. In situations where none of one's pre-existing ends can be satisfied with the few hours on hand, one does better to acquire some new ends quickly—for instance, watching movies

[26] Kolnai (2001) has an interesting discussion of deliberation of ends.

from an unfamiliar genre, or solving a mathematical puzzle that one finds in an inflight magazine. Again, when the environment is supportive, a practically successful person can choose from an extended menu of ends to attain with their available means. The more ends one possesses and is able to acquire, the more likely agent-internal luck is to occur.

4. Conclusion

In this chapter, I argued that free will is compatible with agent-internal luck, on the assumption that rational control is a sufficient control condition for free will. If what I argued is correct, then restrictive compatibilism about agent-internal luck, as proposed by Mele and McKenna, is misguided. My argument also entails that the inclusion of agent-internal luck alone does not constitute a legitimate objection to libertarianism.

Chapter Three

Preserving Resentment for Emotional Intimacy

1. Introduction

A question central to the free will debate concerns whether, if libertarian free will is illusory, resentment should ever be expressed. Some philosophers[27] (call them the *revisionists*) say no, and others[28] (call them the *preservationists*) say yes. Those who favor expressing resentment often justify their position with the moral benefits of doing so. Against those preservationists, I will argue that the moral benefits of expressing resentment are quite limited, citing relevant empirical evidence. Despite the limited moral benefits of expressing resentment, I will argue against the revisionists that we should nonetheless express resentment sometimes for the sake of emotional intimacy, which is an essential part of a flourishing human life. I will propose a version of preservationism—*restricted preservationism*, which recommends limiting the expression of resentment to contexts where emotional intimacy is prized. My position can be considered an effort to reconciliate the two extreme positions of revisionism and standard preservationism, which can be endorsed without assuming that people deserve blame in a basic sense.

In the next section, I will clarify what I mean by resentment and recount the debate between the standard preservationists and the revisionists. Then I will criticize the preservationist argument given by Nichols (2015) and revisionism, respectively in Sections

[27] Examples include Pereboom (2014), Sommers (2007) and Milam (2017).
[28] Examples include P. F. Strawson (1962), Nichols (2015), Wolf (2011), Shabo (2012) and Rei-Dennis (2018).

3 and 4. After that, I will present my position, and respond to potential objections to it in Section 5.

2. Background

Thinking about the free will problem is tricky, in part because we seem to respond to the problem in conflicting ways when we consider it from different perspectives. P. F. Strawson discusses this phenomenon by drawing a distinction between the objective attitude and the reactive attitude. When we adopt the objective attitude, which is to say, when we regard other people and ourselves from an intellectualized, emotionally detached perspective, people appear to lack the kind of free will that makes resentment appropriate. On the other hand, when we are mistreated in everyday contexts, resentment, which Strawson classifies as a reactive attitude, sometimes strikes us as not only psychologically inevitable, but also morally permissible or even morally required.

Not everyone experiences this conflict. Seasoned compatibilists in particular do not find resentment inappropriate for reasons related to free will skepticism, even when they adopt the objective attitude. Nevertheless, this conflict is common for those who at least sometimes find incompatibilism intuitive, which experimental philosophy studies indicate make up the majority of the population[29]. The target audience of this chapter is those who experience this conflict. Accordingly, the question that I'm going to address in this chapter is: Given our conflicting attitudes, should we at least sometimes express our inevitable feelings of resentment?

[29] See Knobe and Nichols (2007).

2.1. Defining resentment

Resentment, as I understand it, is characterized by blaming anger ("anger" thereafter), which is an automatic emotional response to perceived injustice done to oneself. How should we characterize the automatic reaction of anger? Nichols (2015), a preservationist, suggests that the automatic, unreflective action tendency of anger is to retaliate against the offender, which constitutes the narrow psychological profile of anger. Drawing a distinction between anger's narrow, unreflective psychological profile and its wide, reflective psychological profile, Nichols notes that anger can also produce more sophisticated responses in light of reflection. Nichols's model is challenged by Shoemaker (2018), another preservationist, who argues that expression of anger may not involve any intention of retaliation. He writes, "[s]eeing my partner come home drunk and late yet again, I may merely shake my head and quietly shut the bedroom door" (p. 75). Shoemaker proposes that the action tendency of anger be understood as expression of a demand for acknowledgement of one's moral worth.

Shoemaker's criticism of Nichols's account is on target. Anger, even when expressed unreflectively, doesn't always involve an intention to retaliate. However, in suggesting that anger always tends to be expressed as some kind of demand, Shoemaker may have similarly overlooked individual differences in how anger manifests. Anger manifests differently for different individuals and in different contexts[30]. Moreover, there are environmental factors that systematically affect how an individual may unreflectively respond to anger. In the contemporary United States, for instance, an adaptive black man is

[30] Cavell (1979), for example, notes that behaviors associated with anger include "paling or frowning, shaking his fist at another, leaving a room, speaking in a loud voice, pacing nervously, behaving in a generally agitated fashion" (p. 162) and "accuse, chastise, say (intentionally) hurtful words, serve someone right, leave in a rage" (p. 163).

likely slow to express his anger in public, especially when the law enforcement is involved. (In Section 3, I'll discuss how demographic background may affect how one's anger manifests in more detail.)

Anger doesn't seem to exhibit a universally applicable action tendency. Thus, I think that the most immediate automatic response of anger is best characterized by simply the resentful *feeling*, which can be described as combativeness. Accordingly, when resentment is expressed, what gets expressed is this combativeness triggered by perceived injustice, whether or not actual combat is involved. Although the feeling of resentment itself cannot be demonstrated experimentally, it is introspectable.

2.2. The debate

The revisionists argue that resentment presupposes an unjustified belief—that its target is free and thus deserves blame in a basic sense, and is thus morally inappropriate. However, most revisionists acknowledge that we cannot completely eliminate the experience of resentment. What, then, do they think we should do with the ineliminable experience of resentment? The answers from different revisionists vary in detail. For instance, Sommers (2007) argues that we should attempt not to "engage or entertain" the feeling of anger. On the other hand, Pereboom (2014) argues that we should engage resentment in a particular way, by replacing it with sadness or disappointment. Milam (2017) similarly suggests that we should replace anger with other emotions, but provides no principled answers as to what emotions anger should be replaced with. The common position shared by all revisionists is that resentment should always be concealed from other people.

By contrast, preservationists think that resentment should at least sometimes be expressed. While the details differ, the preservationist arguments tend to focus on either (a) the moral benefits of expressing resentment[31], or (b) the inevitability of resentment in intimate relationships[32]. I'll elaborate on the former below, and return to the latter in Sections 4 and 5.

A notable preservationist argument that emphasizes the moral benefits of expressing resentment comes from Nichols (2015), who supports his argument with extensive empirical evidence. Recognizing the tension between our conflicting attitudes towards resentment, Nichols argues that we should express resentment unreflectively, and keep philosophical concerns related to free will skepticism out of our minds in everyday contexts, for the moral benefits of doing so outweigh its costs. According to Nichols's analysis, preserving and expressing resentment has three benefits at the individual level. First, expressing resentment communicates intolerance for abuse. Suppose you have mistreated me. If I express resentment in response to your mistreatment, for example by yelling at you, then you know that I will not tolerate being mistreated again. Second, resentment motivates us to punish against our short-term interests, which serves our long-term interests. This point can be illustrated with an example discussed by Frank (1988), whom Nichols cites. In the example, Smith grows wheat and his neighbor, Jones, raises cattle. Jones can prevent his cattle from eating Smith's wheat by fencing his land, which costs $200. Jones is reluctant to spend the money, but worries that Smith may sue him. Crucially, it would cost Smith $2000 to sue Jones, while any damages to Smith's wheat by Jone's cattle would only cost $1000. Suing Jones would therefore hurt Smith's short-term

[31] Examples include Reis-Dennis (2018), Wolf (2011), Nichols (2015) and Shoemaker (2018).
[32] Examples include Strawson (1962), Shabo (2012).

interests. However, anger would motivate Smith to sue Jones in case of damage after all. The knowledge that Smith may get angry and sue him gives Jones the motivation to fence his land, which serves Smith's long-term interests. Smith's anger thus serves his long-term interests against his short-term interest. This example also illustrates the third benefit of expressing resentment in Nichols's discussion, namely, that it can deter other people from mistreating us[33].

In response to the revisionists, Nichols argues that anger motivates actions in unique ways that some other emotions, such as sadness, do not. Consequently, anger cannot be replaced by sadness as some revisionists, such as Pereboom (2014), suggest. Nichols supports his claim with empirical studies on infants:

> Infants show individual differences in their propensities to feel sad or angry when locked from attaining a desired end—some babies are more likely to feel sad, others to feel angry. Researchers have found that when infants show sadness as their predominant emotion, this is associated with giving up (Lewis and Ramsey 2005, 518), and it seems to be akin to learned helplessness (Abramson et al. 1978). By contrast, infants who respond with anger are more likely to try to overcome an obstacle (Lewis and Ramsey 2005, 518). As a result, sadness seems too behaviorally weak to do the work of anger. (p.162)

At the collective level, Nichols argues that anger promotes social cooperation by motivating altruistic punishment. Nichols recounts the relevant empirical findings in economic games as follows:

[33] Gauthier (1984) similarly advocates expression of our conditional intention as a deterrent policy.

> In public good games… participants are given an allotment of funds and in each round of the game, players are allowed to contribute to the common fund. For every 1 monetary unit an individual contributes, 1.6 units go into the common fund, which is a net benefit for the group, but a net loss for the individual (since he only gets 40 percent of his investment back). Obviously, it's optimal for the group if everyone invests in the common fund, but for each individual, it's selfishly better not to invest. A number of studies have shown that in these games, people will pay to punish players who do not contribute to the common fund (for a review, see Fehr and Fischbacher 2004a)… Why would subjects send their money to punish even when it's obviously not in their material self-interest?... anger is the prevailing explanation… Now, what are the consequences of this anger-driven punishment? As it happens, punishment dramatically changes behavior in these games. Fehr and colleagues have consistently found increased cooperation when punishment is an available option. (pp.158-159)

Despite the impressive empirical evidence cited by Nichols, in the next section, I'll argue that Nichols has exaggerated the moral benefits of expressing resentment.

3. The moral route to preservationism

In this section, I will argue that the moral utility of expressing resentment is quite limited.

3.1. Resentment and individuals

Recall that according to Nichols, expressing resentment has three benefits at the individual level, which are respectively communicative, motivational, and deterrent. These supposed benefits of expressing resentment have been echoed by other preservationists as well. Let us evaluate them one by one.

First, Nichols argues that expressing resentment can communicate intolerance for abuse. In my opinion, there is no obvious reason why intolerance for abuse cannot be communicated with a different emotion, such as sadness, or without any emotions at all. For example, someone can convey intolerance for abuse by simply saying "I won't tolerate this," with a sad face or a Poker face. There is also an extreme tradition of using suicide as an ultimate form of protest. Second, Nichols argues that anger can motivate altruistic punishment while other emotions do not. There is at least some empirical evidence that this claim is not true for adults. In a study conducted by Harle and Sanfey (2007), the subjects were induced to feel sad, amused, or neural with movie clips before they played the Ultimate Game. When compared to the subjects who were induced to feel amused or neutral, the subjects who were induced to feel sad were more likely to reject monetary offers that they found unfair. To be sure, chronic depression can significantly dampen one's motivation even among adults, but we have little evidence to think that momentary sadness lacks the motivational power that resentment has among adults.

While Nichols is correct that resentment sometimes has a unique deterrent effect that other emotions lack, he may have overlooked how the deterrent effect of one's anger is affected by one's social position and demographic background. In Frank's example, Smith's anger can deter mistreatment by Jones partly because Smith has the resource to take legal actions him. What if Smith is too poor to take legal actions against Jones, or what if Smith is

a marginalized minority member? Below, I'm going to recount the relevant empirical evidence that highlights how one's social position and demographic background mediates the deterrent effect of one's anger.

There is evidence that whether expression of anger benefits one depends significantly on one's power position. In a study conducted by Lelieveld, van Dijk, van Beest and van Kleef (2012), participants were asked to play the repeated offer ultimatum bargaining game with an opponent, whose responses were in fact generated by the experimenters. In the game, in the beginning, the participants were asked to choose between two options of distributing 100 chips, with one option favoring the participant and the other option favoring the opponent. Half of the participants believed that they were playing with a high power opponent, whose rejection of a division meant that the participant would receive nothing. The other half of the participants believed that they were playing with a low power opponent, whose rejection of a division only meant that the participant would receive 10% less chips than otherwise. After choosing a division that favored themselves, the participants were manipulated to believe that their opponent expressed anger, disappointment, or no emotion. Later in the game, the participants were asked to propose any division ranging from 0 to 100 chips between themselves and their opponent. When compared to those who received no emotional reactions, participants who received angry reactions offered more chips to their opponent if they believed that they were playing with a high power opponent. However, if the participants believed that they were playing with a low power opponent, they offered fewer chips when they received angry reactions than when they received no emotional reactions. The explanation for this discrepancy seems to be that what reciprocal emotion one's anger triggers in its recipient

depends on one's power position. Lelieveld, et al. found that when anger was believed to come from a high power opponent, it triggered fear in the participants. On the other hand, when anger was believed to come from a low power opponent, it triggered anger in the participants. Since the deterrent effect of anger expression comes from the power of anger to trigger fear, Lelieveld, et al.'s findings suggest that expressing anger is likely to benefit one only when one is in a relatively high power position. This severely limits the utility of anger expression in deterring abuse in everyday life, for a victim of abuse is often not in a higher power position than the abuser; in fact, the reverse is often true. Interestingly, regardless of what power position they believed that their opponent was in, the participants who received disappointment reactions offered more chips to their opponent than when they received no emotional reactions. Moreover, regardless of what power position they believed their opponent was in, the participants experienced guilt when they believed that their opponent expressed disappointment. This suggests that when one's goal is to seek more favorable treatments, expression of disappointment is a more prudent strategy when compared to expression of anger.

Effects of power aside, there is evidence that anger expression backfires when its recipient finds it inappropriate. In a study conducted by van Doorn, van Kleef and van der Pligt (2015), for instance, expressions of anger reduced generous behavior from the participants in a context where they believed that anger was inappropriate. By contrast, expressions of disappointment promoted generous behaviors from the participants in the same context. This again shows that expression of disappointment seems to be a strategy preferable to expression of anger in soliciting cooperative behaviors. At first glance, it seems like findings such as those are irrelevant to Nichols's argument, because the cases of

interest for the preservationists are cases where anger is appropriate. However, notice that how an expression of anger influences another person depends not on how appropriate it is, objectively speaking, but on how appropriate the person perceives it to be from their subjective perspective. In everyday situations, what constitutes a fair arrangement is often ambiguous. Since people tend to interpret ambiguous situations in a manner that favors themselves, what two people think is fair in a situation—and thus whether anger is appropriate in a given situation—is bound to give rise to collisions frequently. This entails that when one expresses one's anger, there is a good chance it is considered inappropriate by its recipient.

So far, my analysis reveals that expressing anger can benefit one only when two conditions obtain—first, when one is perceived to be in a relatively powerful position; and second, when one's expression of anger is perceived to be appropriate. These two conditions often do not obtain, not only because people often lack power and tend to interpret ambiguous situations to favor themselves, but also because whether one is perceived to be powerful and whether one's anger is perceived to be appropriate depends on one's demographic background.

Empirical studies show that what emotion one expresses affects how one is perceived by other people. Tiedens's (2001) experiments showed that when compared to men who expressed sadness in public, those who expressed anger were judged to be more competent, and were assigned a higher social status and a higher salary. In one of the experimental conditions, for instance, the subjects supported Bill Clinton more when they saw him expressing anger about the Monica Lewinsky scandal, versus when they saw him expressing sadness about the scandal. While these findings seem to count in favor of

Nichols's preservationist argument, they cannot be generalized to women. As de Sousa (1990) observes, "[a]n angry man is a 'manly man,' but an angry woman is a 'fury' or a 'bitch.'" (p.259). de Sousa's observation is confirmed by Brescoll and Uhlmann's (2008) findings in their follow-up study on Tieden (2000). In one of Brescoll and Uhlmann's experiments, for instance, the subjects were shown a videotaped job interview in which (a) either a male interviewee or a female interviewee described feeling (b) either sad or angry about an identical incident. The findings are stunning. In the male interviewee conditions, the angry interviewee was judged to deserve a higher social status, a higher salary, and perceived to be more competent than the sad interviewee. In the female interviewee conditions, however, the effects of anger were reversed. In the female interviewee conditions, the angry interviewee was judged to deserve a lower social status, a lower salary, and perceived to be less competent than the sad interviewee. Furthermore, Brescoll and Uhlmann found that when anger was expressed by a woman, it was significantly more likely to be judged as "out of control" (p. 271). In other words, when a woman expresses anger, not only is she perceived to be less powerful than her male counterparts, but her anger is also less likely to be considered as an appropriate reaction to genuine mistreatments. As a result, expressing anger is less likely to be beneficial in ways that Nichols suggests if one is a woman. Brescoll and Uhlmann's findings also highlight additional costs of expressing anger that Nichols does not consider. Since it is among most people's top concerns to receive a good salary and be perceived as competent in the workplace, expressions of anger in public come with significant costs that most women cannot easily discount. Because the narrow psychological profile of one's anger is shaped

by the environment, the unreflective tendency of a well-adjusted woman's anger is likely non-expression.

At least in the United States, how one's expression of anger is perceived depends on one's racial background as well. This is particularly true of African Americans, who have prudential reasons to suppress their anger in public. The following story told by an African American student, recorded in Solorzano and colleagues's (2000) campus study, illustrates the relevant psychological complications:

> it's 11 o'clock [at night] and all of a sudden, [campus police] sweeps up… all here for us who are not displaying any type of violence or anything like that… but we're upset. And we're saying at the same time, we're feeling restricted because if we act in a way we want to react—number one, we're going to go to jail; number two—it's just going to feed into the stereotype that we're supposed to be violent... (p.69, emphasis added)

As in the case of women, expressions of anger from an African American are less likely to be considered appropriate. Moreover, the potential costs of expressing anger are even higher for African Americans than for (non-African American) women. When expression of anger backfires for an African American in public, it may result in legal trouble and, in extreme cases, live fire. Thus, the unreflective action tendency of anger for a well-adjusted African American is also likely to be non-expression.

Even if what I've argued so far is correct, there remains the possibility that expressing anger is beneficial for those with certain demographic backgrounds—i.e., Caucasian men in relatively powerful positions. To this I have two responses. First, while

anger expression is in general less costly when it comes from members of this demographic group, its benefits are limited, for the reason that they have fewer occasions for it. Power suffices to deter mistreatment; anger expression is unnecessary. After all, there is a reason why there is a nationwide "Black Lives Matter" movement but not a "White Lives Matter" movement, and a "Me Too" movement for women but not an equivalent movement for men; members of minority groups are much more likely to be victims of injustice. Second, power is relative and context specific, and thus no one is always in a relatively powerful position. Someone who is powerful at work may have little they can do when it comes to influencing their family members, for instance.

At this point, one may agree that the effects of expressing one's anger are mediated by one's social position and demographic background in non-intimate relationships, but maintain that such mediating effects are absent in intimate relationships and thus, everyone can reap the moral benefits of unreflective anger expression in intimate relationships. My response consists of three points. First, power dynamics exist in intimate relationships as well. This is most obviously true for parental relationships, but it is also common to encounter relationships between adults where one person's preference tends to dominate another person's preference, whatever the explanation may be. In situations like this, the anger of one person is usually given more weight than that of the other person in the relationship. Second, anger can deter mistreatment only in relatively healthy relationships, but is useless in truly abusive relationships. Third, our emotional responses are often biased in intimate relationships. I'll elaborate the last two points further.

In truly abusive relationships, a victim's anger often has little deterrent effect. If anger is a more useful response to abuse, we would expect to see victims of domestic abuse

being given trainings to express their anger more effectively. But we do not. In fact, if a victim of domestic abuse chooses to stay in an abusive relationship and express anger to their abuser, instead of moving on from the relationship, it is considered a sign of battered spouse syndrome. That is because in a truly abusive relationship, anger from the victim is useless. On the flip side, when one's anger is useful in a relationship, the relationship is relatively healthy; what triggers anger is often morally trivial, or even morally unproblematic. Consider what some preservationists cite as examples of wrongdoing in intimate relationship. Wolf (2011) complains of her husband's "tendency to say, 'I'm ready to go' only then to keep… (her) waiting while he spends an extra five minutes finding his glasses, washing his coffee cup, and getting his books together" (p.334). While Wolf's anger is understandable, whether her anger can change her husband's behavior is, in my opinion, of little moral consequence. It is a common observation that wrongdoings in a healthy relationship are often instances like that. Other preservationists' examples highlight how, in intimate relationships, anger is sometimes triggered by possessiveness rather than genuine wrongdoing. Consider Shoemaker's (2018) example of seeing his partner "come home drunk and late" (p. 75). It is unclear precisely what the wrongdoing here is. After all, what *moral* reasons does an adult have to not go home drunk and late as much as she wants? While infidelity may be implied in the example, the ethical status of infidelity is itself morally questionable. In intimate relationships, we often see people getting angry in situations like this out of jealousy, rather than perception of genuine injustice. Reis-Dennis's (2018) example of being angry in response to a relationship not remaining as it was similarly highlights how anger in intimate relationships is often triggered by unjustified possessiveness. When anger is triggered by something morally trivial or

unproblematic, the moral benefits of expressing it are also limited. Besides, if a relationship is healthy enough for one's expression of anger to be useful for behavioral modification, one's expression of sadness is also useful. After all, those we are in a healthy relationship with are intrinsically motivated to avoid seeing us sad. That is to say, when one's expression of anger is useful in a relationship, it is, paradoxically, unnecessary.

Our automatic anger response is systematically biased. For instance, how angry we feel about an action depends not only on the intention of the actor, but also on consequences of the action over which the actor has no control[34]. Our automatic anger response also exhibits cultural bias; how angry one feels about an action depends partly on what culture one grows up in. In a sexist culture, for instance, a woman not covering herself up according to religious standards is enough to trigger anger. It is safe to assume that no one's emotional system is immune from cultural biases. Our emotional responses are especially biased when it comes to those we like. It is much easier to be objective about someone to whom we are emotionally indifferent than someone to whom we are emotionally vulnerable. We find it easy to overlook the shortcomings of those we like due to idealization, yet we resent them when they do not live up to our idealization. While we can arguably be more objective about our resentment, who wants someone who scrutinizes our everyday actions like an impartial judge as a lover, or even as a friend? Besides, to adopt the objective attitude also means to consider concerns related to free will skepticism relevant. From that perspective, resentment appears inappropriate regardless of what triggers it, if one is inclined to find free will skepticism intuitive.

[34] See Kahneman and Federick (2002) for discussions of outrage heuristics.

3.2. Resentment and society

According to Nichols, unreflective anger expression has moral benefits not only for the individuals, but also for society. Suppose we compare unreflective anger expression to punishment from an informal criminal system; we may ask, under what conditions can this system improve the society morally? One such condition is that people's unreflective anger responses reliably track the degree of wrongdoing. But as I pointed out above, our automatic anger responses are systematically biased. Resentment can be triggered by genuine wrongdoing, but it can also be triggered by violation of unjust norms or just personal prejudice. Another condition under which unreflective anger expression can benefit the society as a whole is that the same wrongdoing is punished equally regardless of who the victim is. If all members of the society can deter mistreatment equally effectively with their anger, perhaps different people's biases can cancel each other out. My analysis above reveals that this is not the case. A likely consequence of having only a fragment of the population express their anger unreflectively is that anger is used to enforce oppressive, racist, sexist norms that uniquely favor them. Arguably this is already the reality.

4. Revisionism

Although the moral benefits of expressing resentment are limited, I do not think that resentment should therefore always be concealed. In this section, I'll criticize revisionism. In a nutshell, I will argue that resentment should sometimes be expressed because it is an authentic human emotion, and therefore expression of it enables emotional intimacy.

A recurring claim among the revisionists is that expression of resentment is unnecessary for good interpersonal relationships, and thus expression of resentment can be replaced by expression of other emotions. Pereboom (2014), for instance, argues that expression of sadness or disappointment can communicate "the relevant information" (p. 146) that we would like to communicate via expression of resentment.

Some preservationists object to the revisionist proposal by arguing that expression of resentment is inevitable in intimate relationships. Shabo (2012), for instance, writes that "at least in the context of reciprocal love, our caring in an essentially personal way about the other's feelings and behavior toward us ensures that most of us cannot reliably forgo or disavow feelings of resentment in response to significant violations of our demand for due regard…" (p. 113). My response to revisionism differs from this line of reasoning in two ways. First, I think that expressing resentment in intimate relationships is not only inevitable, but also valuable, such that we should sometimes express our resentment even if doing so is psychologically avoidable. Second, I don't think that expressing resentment is valuable only when it is a response to some significant moral violation. On the contrary, I think that expression of resentment can be valuable even when it is morally inappropriate.

Intimacy varies along different dimensions; not all of them are equally important. One type of intimacy that is necessary for a flourishing human life is emotional intimacy, which empirical studies indicate is important for one's physical and mental health. Sinclair and Dowdy (2005), for instance, found that emotional intimacy is positively correlated with "social support, self-efficacy, perceived health competence, reappraised coping behaviors, life satisfaction, and positive effect" (p. 193) and negatively correlated with "perceived stress levels, helplessness, negative pain coping behaviors, pain and fatigue" (p. 193).

Essential to emotional intimacy is disclosure of how one feels[35]. As such, emotional intimacy is a form of honesty about what one is like. By concealing resentment, one is dishonest about how one really feels. Expressing sadness or disappointment when one is feeling resentment does not communicate an important piece of information—namely, how one really feels, which is combativeness. Revisionism is therefore a form of emotional deception that limits the possibility of emotional intimacy.

Emotional intimacy can be achieved not only through disclosure of morally appropriate emotions, but also through that of morally inappropriate emotions. We often see good friends bonding over their shared amusement about a politically incorrect joke. Pet owners often bond over sharing their inappropriate resentment in response to their pet's destructive behavior. While there are legitimate concerns regarding expression of morally inappropriate emotions, notice that in the context of intimate relationships, norms, if they exist at all, are recalibrated. Consequently, norms that govern what kinds of expressions are considered appropriate are also recalibrated. This is most obvious in the romantic context, but it also happens between good friends. The recalibrated outcome differs in each relationship, but usually, some expression of inappropriate attitudes is not only tolerated, but also expected or even desired. When someone always acts perfectly appropriately around us, we sense emotional distance from them.

Madonna accused Lady Gaga of copying her music, which Lady Gaga denied. When asked to comment on Madonna's accusation, Lady Gaga said that she wanted Madonna to push her up against a wall and tell her that she was "a piece of shit." [36] Lady Gaga's comment would be puzzling if not for the understandable appeal of emotional intimacy—

[35] See Laurenceau, Pietromonaco and Barrett. (1998).
[36] In Campbell, Moukarbel, Parry, Gaga, & Moukarbel (2017).

Why else would she want to be treated aggressively for what she thinks is an unfair judgement? But emotional intimacy with a person may entail experiencing that person's feeling for what it is, whether or not the feeling is morally appropriate.

In suggesting that expression of resentment can be adequately replaced by that of sadness or disappointment, Pereboom (2014) cites parental relationships as an example, which he thinks can be generalized to other forms of intimate relationships. We can now see why this generalization is problematic. While parental relationships are intimate in many ways, on the assumption that the child is not an adult, emotional intimacy should be limited in this context. Indeed, inappropriate emotions from a parent should be concealed from a child, for a child does not have the capacity to properly process inappropriate emotions from adults, and can thus be traumatized by them.

Even if we are willing to forego emotional intimacy with other people, it is difficult to conceal resentment from someone who knows us very well. Pereboom argues that we can avoid revealing our resentment by citing the example of trainers of dangerous animals, who must avoid provoking the animals by revealing any anger. But we should hope that we have at least someone who knows us better than the dangerous animals know their trainers. While dangerous animals can only assess their trainer's emotional state through their non-verbal expressions, people can assess each other's emotional state through verbal expressions as well. To someone who knows us very well, even silence may indicate our resentment. Thus, unless we don't have someone who knows us very well around, the only way to completely conceal our resentment from other people is to conceal it from ourselves. Doing so, however, compromises our emotional intimacy with ourselves, which is arguably the most important form of emotional intimacy. As the Oracle at Delphi had it,

"know thyself." Self-knowledge includes knowledge of one's feelings, regardless of whether they are appropriate. Ignorance of how one feels has bad consequences. According to the Freudian tradition, chronic depression is anger turned inward, which results from repressing the knowledge of how one feels. If the Freudian insight is correct, endorsement of revisionism is a risk factor of chronic depression.

5. The non-moral route to preservationism

Thus far, I have argued for two claims. First, expressing resentment has little moral utility. That is to say, in most cases, whether or not we express our resentment makes little moral difference to the world. Second, a complete prohibition of expressing resentment limits the possibility of emotional intimacy, which is essential for human flourishing. A natural suggestion follows: that expression of resentment be restricted to contexts where emotional intimacy is prized. Call this position restricted preservationism.

Despite what the word "intimacy" may suggest, the value of emotional intimacy is not exclusive to relationships that are otherwise intimate. When Barack Obama shared his anger over climate change issues in public, for instance, he achieved emotional intimacy with people he had never encountered. A dose of emotional intimacy can also make transactional relationships seem less transactional. When emotional intimacy is prized is a matter of personal value. As I discussed above, expressing resentment can be costly with little moral utility for members of most demographic groups. But good things in life are often costly.

Emotional intimacy comes in degrees, which is reflected in different forms of expressions of resentment. With a high degree of emotional intimacy, expression of

resentment can be barbaric and aggressive. On most occasions, however, with a more limited degree of emotional intimacy, I recommend that resentment be expressed in a civilized, non-violent manner. Gandhi's civil disobedience movement is an example of civilized expression of resentment. Other examples include expression through art, and good old passive-aggression. Sometimes, deciding how to express one's resentment to a person is a part of deciding what one's relationship with that person is like. However, a high degree of emotional intimacy may require one to refrain from too much deliberation before sharing an emotion. A relationship that involves much calculation may have the appearance of intimacy, but is not that intimate after all. In the literature, a free choice is sometimes thought of as requiring some level of reflection (e.g., Lehrer (draft), Frankfurt (2003), Watson (2003)). An interesting implication of my analysis is that a high degree of emotional intimacy sometimes precludes one's choices from being free on those accounts.

There are cultural as well as gender differences when it comes to emotional intimacy. For instance, women tend to enjoy more emotional intimacy than men, while Westerners tend to enjoy more emotional intimacy than Asians. This should be taken into account when deciding what manner of expressing resentment may be appropriate in a context.

Objections to my position may come from two opposite directions. On one hand, a revisionist may argue that I allow too much expression of resentment. After all, expressing resentment can be harmful. On the other hand, a preservationist may argue that I don't allow enough in the way of expressing resentment. I'll address each objection below.

To the objection that my position allows too much expression of resentment, I have two responses. First, emotional intimacy is essential for human flourishing, as I argued

above. If any harm is done due to expressing resentment, it is a necessary evil. Second, rationality can restrain expressing resentment in contexts where emotional intimacy should be limited (such as in the context of parental relationships).

Let us move on to the objection that my position allows too little expressing resentment. The objection may come in two forms. First, the objection may be that by restricting expressing resentment to only contexts where emotional intimacy is prized, we cannot effectively deter mistreatments by other people. Second, the objection may be that by restricting expressing resentment to only situations where emotional intimacy is prized, what I propose is a form of affective injustice.

Regarding the objection that we cannot effectively deter mistreatments by other people if we restrict expression of resentment, I argued above in Section 3 that the deterrent effects of expressing resentment are conditional and thus limited. Here, I would like to add that we don't live in the state of nature, where expressing resentment is arguably more important for deterrent purposes; we live in a society with protection by the law. If we are seriously mistreated, we can call the police. While the legal system is imperfect, it is a more powerful tool than most people's anger in deterring mistreatment.

The objection concerning affective injustice is adapted from Srinivasan (2018), who coins the term and argues that prohibition of apt but counterproductive anger is a form of affective injustice. Crucial to her argument is that apt anger, as a means to affectively appreciating the injustice of the world, has an intrinsic value that cannot be reduced to its instrumental value. Thus, even when anger has little moral utility, as I argue, it remains intrinsically valuable when it is apt. To this I have two responses. First, to those who are inclined to find free will skepticism intuitive, anger never seems morally appropriate.

Second, affectively appreciating the world is a form of emotional intimacy with oneself, which discloses to oneself how one feels about the world. One's value system may be such that one always prizes emotional intimacy in contexts where one's anger is apt in Srinivasan's sense—when it reflects the degree of moral violation in the world. However, I have reservations about the possibility of anyone's anger system being apt that way. After all, given the numerous occasions for anger that crop up around the world every day, an anger system that reflects objective moral violation in the world would be hyperactive and thus unsustainable.

6. Conclusion

Expressing resentment has little moral utility, and is often costly. Nevertheless, we should sometimes express our resentment for the sake of emotional intimacy, which is a form of emotional honesty. Emotional intimacy is valuable because it reveals attitudes that are ours; it unites those who speak the common language of human emotions, which are sometimes appropriate, and sometimes not.

P. F. Strawson argues that in seeking a moral justification for our reactive attitudes, philosophers overintellectualize them; they wear a "pitiful intellectualist trinket" (p.92) of fittingness against the recognition of their humanity. Interestingly, though many philosophers consider themselves to be Strawsonian, most of them are reluctant to take off the trinkets. My discussion is a revival of Strawson's insight that reactive attitudes should be affirmed, not primarily because of their moral utility, but simply because they are parts of our humanity.

References

Amabile, T. M. (1979). Effects of external evaluation on artistic creativity. *Journal of Personality and Social Psychology, 37*(2), 221–233

Amabile, T., Goldfarb, P. & Brackfleld, S. (2009). Social influences on creativity: Evaluation, coaction, and surveillance. *Creativity Research Journal, 3*(1), 6-21.

Ayer, A. J. (1972). *Philosophical essays*. London: Palgrave Macmillan.

Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning. *Organizational Behavior and Human Decision Processes, 42,* 88-110.

Binmore, K. (2007). *Game theory: A very short introduction.* New York: Oxford University Press.

Brescoll, V., & Uhlmann, E. (2008). Can an angry woman get ahead? *Psychological Science, 19,* 268-275.

Buss, S. (2020). Some reflections on the relationship between reason and the will. In R. Chang & K. Sylvan (Eds.), *The Routledge handbook of practical reason* (pp. 196-213). New York: Routledge.

Campbell, B., Moukarbel, C., Parry, H., Gaga, L. (Producers), & Moukarbel, C. (Director). (2017). *Gaga: five foot two* [Motion picture]. United States: Live Nation Productions.

Cavell, S. (1979). *The claim of reason: Wittgenstein, skepticism, morality, and tragedy.* New York: Oxford University Press.

Chalmers, D. J. (2011). *Philosophical Review, 120*(4), 515-566.

Chan, H., Deutsch, M., & Nichols, S. (2016). Free will and experimental philosophy. In J. Sytsma & W. Buckwalter (Eds.), *A companion to experimental philosophy* (pp. 158-172). Malden, MA: Wiley Blackwell.

de Calleja, M. (2014). Cross-world luck at the time of decision is a problem for compatibilists. *Philosophical Explorations, 17,* 112-25.

Dennett, D. C. (2001). In Darwin's wake, where am I? *Proceedings and Addresses of the American Philosophical Association, 75*(2), 13-30.

de Sousa, R. (1990). *The rationality of emotion.* Cambridge, MA: The MIT Press.

Gauthier, D. (1984). Deterrence, maximization, and rationality. *Ethics, 94*(3), 474-495.

Fischer, J. (2006). *My way: Essays on moral responsibility.* Oxford: Oxford University Press.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: An essay on moral responsibility.* Cambridge: Cambridge University Press.

Frankfurt, H. (2003). Freedom of the will and the concept of a person. In G. Watson (Ed.), *Free will* (pp. 322-336). New York: Oxford University Press.

Frank, R. (1988). *Passion within reason.* New York: W. H. Norton.

Harle, K. M., & Sanfey, A. (2007). Incidental sadness: Biased social economic decisions in the ultimatum game. *Emotion, 7,* 876-887.

Hennessey, B.A. (2001). The social psychology of creativity: Effects of evaluation on intrinsic motivation and creativity of performance. In S.G. Harkins (Ed.), *Multiple perspectives on the effects of evaluation on performance* (pp. 47-76). Boston, MA: Springer.

Hume, D. (2006). An enquiry concerning the principles of morals. In S. M. Geoffrey (Ed.), *Moral philosophy* (pp. 185-310). Indianapolis, IN: Hackett.

Hume, D. (2011). An enquiry concerning human understanding. In T. Griffith (Ed.), *Hume: The essential philosophical works* (pp. 573-706). Hertfordshire: Wordsworth.

Ismael, J. (2016). *How physics makes us free.* Oxford: Oxford University Press.

Kahneman, D., & Federick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). Cambridge: Cambridge University Press.

Kane, R. (1996). *The significance of free will.* Oxford: Oxford University Press.

Kane, R. (2005). *A contemporary introduction to free will.* Oxford: Oxford University Press.

Kant, I. (1788). *Critique of pure reason.* Cambridge: Cambridge University Press.

Knobe, J., & Nichols, S. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous, 41,* 663 – 685.

Kolnai. (2001). Deliberation is of ends. In E. Millgram (Ed.), *Varieties of practical reasoning* (pp. 259-278). Cambridge, MA: The MIT Press.

Laurenceau, J., Pietromonaco, P., & Barrett, L. (1998). Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of Personality and Social Psychology, 74*(5), 1238-1251.

Lehrer, K. (draft). *Ultimate freedom.*

Lehrer, K. (2012). *Art, self and knowledge*. New York: Oxford University Press.

Lelieveld, G., van Dijk, E., van Beest, I., & van Kleef, G. (2012). Why anger and disappointment affect other's bargaining behavior differently: The moderating role of power and the mediating role of reciprocal and complementary emotions. *Personality and Social Psychology Bulletin, 38*(9), 1209-1221.

Lenman, J. (2000). Consequentialism and cluelessness. *Philosophy and Public Affairs, 29,* 342-370.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility.* Oxford: Oxford University Press.

List, C. (2014). Free Will, determinism, and the possibility of doing otherwise. *Nous, 48*, 156-178.

McKenna, M. (draft). Facing the luck problem for compatibilists.

McKenna, M. (2014). Resisting the manipulation argument: A hard-liner takes it on the chin. *Philosophy and Phenomenological Research, 89,* 467-84.

McKenna, M. (2013). Reasons-responsiveness, agents, and mechanisms. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility (Vol. 1)* (pp. 151-84). Oxford: Oxford University Press.

McKenna, M. (2008). Ultimacy and sweet Jane. In N. Trakakis & D. Cohen (Eds.), *Essays on free will and moral responsibility* (pp. 186-208). Newcastle: Cambridge Scholars Publishing.

McKenna, M., & Pereboom, D. (2016). *Free will: A contemporary introduction.* New York: Routledge.

Mele, A. (2006). *Free will and luck.* Oxford: Oxford University Press.

Milam, P. (2017). In defense of non-reactive attitudes. *Philosophical Explorations, 20*(3), 294-307.

Mill, J. S. (2006). On genius. In J. Robson & J. Stillinger (Eds.), *Collected works of John Stuart Mill (Vol. 1)* (pp. 327-340). Buffalo, NY: Liberty Fund.

Millgram, E. (2005). *Ethics done right.* New York: Cambridge University Press.

Nelkin, D. (2011). *Making sense of freedom and responsibility*. Oxford: Oxford University Press.

Nichols, S. (2015). *Bound: Essays on free will and responsibility.* Oxford: Oxford University Press.

Nichols, S., Pinillos, A., & Mallon, R. (2016). Ambiguous reference. *Mind, 125*, 145-175.

Nietzsche, F. (1974). *The gay science.* New York: Vintage Books.

Nozick, R. (1968). *Philosophical explanations*. Cambridge, MA: Harvard University Press.

O'Connor, T. (2000). *Persons and causes.* New York: Oxford University Press.

Paul, L. A. (2014). *Transformative experience.* Oxford: Oxford University Press.

Pereboom, D. (2014). *Free will, agency and meaning in life.* New York: Oxford University Press.

Rei-Dennis, S. (2018). Anger: scary good. *Australian Journal of Philosophy, 97,* 451-464.

Sartorio, C. (2021). Indeterministic compatibilism. In M. Hausmann and J. Noller (Eds.), *Free will: Historical and analytic perspectives* (pp. 205-27). London: Palgrave Macmillan.

Sartorio, C. (2016). *Causation and free will.* Oxford: Oxford University Press.

Sartre, J. P. (1992). *Being and nothingness.* New York: Washington Square Press.

Scanlon, T. M. (2013). Giving desert its due. *Philosophical Explorations, 16,* 101-116.

Schechtman, M. (2014). *Staying alive: Personal identity, practical concerns, and the unity of a life*. Oxford: Oxford University Press.

Schmidtz, D. (2002). How to deserve. *Political Theory, 30,* 774-799.

Schmidtz, D. (1994). Choosing ends. *Ethics, 104,* 226-251.

Shabo, S. (2012). Incompatibilism and personal relationships: Another look at Strawson's objective attitude. *Australian Journal of Philosophy, 90*(1), 131-147.

Shalley, C. E. (1995). Effects of coaction, expected evaluation, and goal setting on creativity and productivity. *Academy of Management Journal, 38*(2), 483-503.

Shoemaker, D. (2018). You oughta know: Defending angry blame. In M. Cherry & O. Flanagan (Eds.), *The moral psychology of anger* (pp. 67–88). Lanham, MD: Rowman & Littlefield.

Simon, H. (1997). *Models of bounded rationality.* Cambridge, MA: The MIT Press.

Sinclair, V., & Dowdy. S. (2005). Development and validation of the emotional intimacy scale. *Journal of Nursing Measurement, 13*(3), 193-206.

Singer, P. (2016). *Famine, affluence, and morality.* Oxford: Oxford University Press.

Sivanathan, N., & Kakkar, H. (2017). The unintended consequences of argument dilution in direct-to-consumer drug advertisements. *Nature Human Behaviour, 1,* 197-802.

Skinner, B. F. (1948) *Walden two.* Indianapolis, IN: Hackett Publishing Company.

Slote. (2001). Moderation and satisficing. In E. Millgram (Ed.), *Varieties of practical reasoning* (pp. 221-236). Cambridge, MA: The MIT Press.

Smilansky, S. (2000). *Free will and illusion.* Oxford: Clarendon Press.

Smilansky, S. (2003). Compatibilism: the argument from shallowness. *Philosophical Studies, 115,* 257-82.

Solorzano, D., Ceja, M., & Yosso, T. (2000). Critical race theory, racial microaggressions, and campus racial climate: The experiences of African American college students. *Journal of Negro Education, 69,* 60-73.

Sommers, T. (2007). The objective attitude. *Philosophical Quarterly, 57*(228), 1-21.

Srinivasan, A. (2018). The aptness of anger. *The Journal of Political Philosophy, 26*(2), 123-144.

Strawson, G. (2004). Against narrativity. *Ratio, 17*(4), 428-452.

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies, 75,* 5-24.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy, 48,* 1-25.

Tieden, L. (2000). Anger and advancement versus sadness and subjugation: The effect of negative emotion expressions on social status conferral. *Journal of Personality and Social Psychology, 80*(1), 86-94.

van Doorn, E., van Kleef, G., & van der Pligt, J. (2015). How emotional expressions shape prosocial behavior: Interpersonal effects of anger and disappointment on compliance with requests. *Motivation and Emotion, 39,* 128–141.

Vargas, M. (2012). Why the luck problem isn't. *Philosophical Issues, 22,* 419-436.

Vargas, M. (2007). Revisionism. In J. Fischer, R. Kane, D. Pereboom & M. Vargas (Eds.), *Four views on free will* (pp. 126-165). Oxford: Blackwell Publishers.

Watson, G. (2003). Free agency. In G. Watson (Ed.), *Free will* (pp. 337-351). New York: Oxford University Press.

Wolf, S. (2011). Blame, Italian style. In R. Wallace, R. Kumar & S. Freeman (Eds.), *Reasons and recognition: Essays on the philosophy of T.M. Scanlon* (pp. 332-347). New York: Oxford University Press.

Wolf, S. (1990). *Freedom within reason.* Oxford: Oxford University Press.