Carving a Life from Legacy:

Frankfurt's Account of Free Will and Manipulation

in Greg Egan's "Reasons to Be Cheerful"

Taylor W. Cyr

University of California, Riverside

Abstract

Many find it intuitive that having been manipulated undermines a person's free will. Some have objected to accounts of free will like Harry Frankfurt's (according to which free will depends only on an agent's psychological structure at the time of action) by arguing that it is possible for manipulated agents, who are intuitively unfree, to satisfy Frankfurt's allegedly sufficient conditions for freedom. Drawing resources from Greg Egan's "Reasons to Be Cheerful" as well as from stories of psychologically sophisticated artificial intelligence (such as Isaac Asimov's "The Bicentennial Man"), I rebut this objection to "structuralist" accounts of free will, arguing that the very possibility of free will for persons like us requires that we admit that a person can be free even when lacking control over the character from which she acts. I conclude with some implications for the freedom and personhood of artificial intelligences.

1. Introduction

One feature of persons—arguably a *distinctive* feature of persons—is their capacity for acting with freedom of the will. According to Harry Frankfurt's (1971) influential account, acting from a will that is free is a matter of having a certain sort of psychological structure at the time of action, specifically one in which a person not only has the desire to perform an action but also has a second desire that the first desire be effective in leading to action. Since Frankfurt's account says acting with free will is a matter of having a certain sort of psychological structure at the time of action, the account is a "structuralist" one.[1] A popular objection to Frankfurt's account (and to structuralist accounts more generally) is that it is possible for the account's alleged sufficient conditions on free will to be satisfied by agents who have been manipulated and who thus (because of the manipulation) appear unfree. For example, if we were to discover that a person who committed some heinous deed had recently been *brainwashed* into thinking that the heinous deed was the best course of action, many

Journal of Science Fiction and Philosophy vol.1 (2018)

of us would be disposed to say that this person was not free *even if* this agent had, say, a desire that the desire to commit the heinous deed be effective. It would seem, then, that whether one acts with free will (and thus satisfies an important condition on being a person) is partly a matter of how one came to perform the action in question, not merely a matter of having a certain psychological structure at the time of action.

While not a case of manipulation *per se*, Greg Egan creates a relevantly similar scenario in his short story, "Reasons to Be Cheerful" (1998). Late in the story, Mark (the narrator) undergoes a procedure that results in a psychological profile (including desires and preferences) that is influenced not by his own past choices, character, and preferences, but rather by the amalgamation of four thousand contributing networks (the psychological profiles of four thousand deceased human beings). Not long after receiving this psychological profile, Mark uses the technology he's been given to choose to fall in love with another character, Julia, and he does fall in love with her. At the time of choosing Julia, Mark satisfies the conditions on Frankfurt's account of free will, yet Mark expresses the very worry that philosophers have raised for accounts like Frankfurt's, namely that, in his case, it isn't really *him* that chose Julia, given the influence of others' psychological profiles on his own. Later on, however, Mark also explores a type of response to the worry about such external influence on his behavior; while his case may seem extraordinary, in relevant respects he is no different from an ordinary person, who must "carve a life out of the same legacy," namely the legacy of natural selection, parental and societal influence, and other shaping factors.

In this essay, I will develop this latter suggestion of Mark's into a full-fledged response to the manipulation objection to accounts of free will like Frankfurt's. I will argue that, although cases of manipulation (and scenarios like Mark's) may appear at first glance to preclude freedom, any account according to which we may become free persons must allow that we can act with free will even when acting from psychological profiles over which we had no control. More specifically, since each of us (ordinary human beings) began to exist, and since we lacked control over our psychological profiles at the start of our careers as agents, we are not relevantly different (at certain parts of our lives) from manipulated agents (or from agents like Mark). And insofar as we take our capacity to act with freedom of the will to be a centrally important feature of persons, it is crucial that, in order to have an adequate account of free will, we must allow for freedom in cases like Mark's.

I will proceed as follows. In the next section, section 2, I will summarize the details of Frankfurt's theory of freedom. Then, in section 3, I will fill out the objection to Frankfurt's account that I mentioned above, namely the threat from manipulation. In section 4, I will provide a brief recap of the plot of Egan's "Reasons to Be Cheerful" and will discuss the more general theme, pervasive in science fiction, of the freedom (or lack of freedom) enjoyed by artificial intelligences that are designed to satisfy conditions like those of Frankfurt's theory of freedom. Finally, in section 5, drawing resources from the examples from science fiction, I will argue that the threat from

manipulation fails to undermine accounts of freedom like Frankfurt's. I conclude by briefly summarizing my response to that objection and by considering the implications for the freedom and personhood of artificial intelligences.

2. A Theory of Freedom (and Personhood)

According to Frankfurt, "one essential difference between persons and other creatures is to be found in the structure of a person's will" (1971, 6).[2] While many creatures have wants and desires to act (or to omit to act) in various ways, Frankfurt thinks that *persons* have the capacity to want or desire to have (or not to have) certain wants or desires to act in various ways. Desires of the first sort are "first-order desires," and these take such things as courses of action as their object. A person's "will," according to Frankfurt is to be identified with certain first-order desires, namely those that are effective in bringing about a person's action. Desires of the second type, "second-order desires," take first-order desires as their object, either by desiring to have a certain first-order desire or by desiring that a certain first-order desire bring about an action. For example, although a person may lack the first-order desire to exercise, she may nevertheless have the second-order desire *that she have the desire to exercise*. For another example, a person may have conflicting first-order desires, such as the desire to exercise now and the desire to relax now, and she may have the second order desire *that the desire to exercise be effective*. In Frankfurt's view, this capacity for second-order desires about which first-order desires are effective is a distinctive feature of persons, and we can understand what it is for a person to act with freedom of the will by reference to the structure of the person's will.

What is it, then, on Frankfurt's account, for a person to act with freedom of the will? Frankfurt offers the following summary of his view:

> A person's will is free only if he is free to have the will he wants. This means that, with regard to any of his first-order desires, he is free either to make that desire his will or to make some other first-order desire his will instead. Whatever his will, then, the will of the person whose will is free could have been otherwise; he could have done otherwise than to constitute his will as he did. (Frankfurt 1971, 18–19)

As long as a person is free in choosing which of her first-order desires will lead her to action, the person acts with free will when the selected first-order desire brings about the action. So, on Frankfurt's account, if I desire that my desire for coffee lead me to the action of ordering coffee (rather than, say, my desire for tea leading me to the incompatible action of ordering tea), and if my desire for coffee is effective in bringing me to ordering some, then I count as acting from free will in ordering my coffee. The conditions of Frankfurt's account pertain only to the internal structure of an agent, and the conditions are sufficient, Frankfurt thinks, for acting with free will.

One especially important feature of Frankfurt's theory of freedom (a feature shared by all structuralist accounts) is that a person's history makes no difference as to whether the person acts with free will. Instead, on this type of account, all that matters for freedom is the person's psychological structure at the time of action. This feature of the account gives rise to the worry that certain persons who intuitively lack freedom may nevertheless satisfy Frankfurt's sufficient conditions on acting from free will. Let us turn now to this objection.

3. The Threat from Manipulation

As we have already seen, Frankfurt's sufficient conditions on free will make no reference to how a person comes to have the psychological structure specified by the account's conditions. An implication of this, which Frankfurt himself notes, is that "some agency other than a person's own [may be] responsible (even *morally* responsible) for the fact that he enjoys or fails to enjoy freedom of the will" (1971, 20). But a popular worry for views like Frankfurt's arises because of this implication.

To see the worry, consider the following story of covert manipulation, adapted from stories told in various places by Alfred Mele (see, e.g., Mele 2016, 72–74):

> Beth is an exceptionally sweet person, though she was not always that way. When she was in high school, she came to view herself, with some justification, as self-centered, petty, and somewhat cruel. She worked hard to improve her character over the course of several years, and she succeeded.
>
> Chuck, by contrast, enjoyed torturing animals as a young boy, but he was not wholeheartedly behind this. These activities sometimes caused him to feel guilty, he experienced bouts of squeamishness, and he occasionally considered abandoning animal torture. However, Chuck valued being the sort of person who does as he pleases and who unambivalently rejects conventional morality as a system designed for and by weaklings. He set out to ensure that he would be wholeheartedly behind his torturing of animals and related activities, including his merciless bullying of vulnerable people, and he succeeded.
>
> When Beth crawled into bed last night she was an exceptionally sweet person, but she awakes with a desire to stalk and kill a neighbor, George. Although she had always found George unpleasant, she is very surprised by this desire. What happened is that, while Beth slept, a team of psychologists that had discovered the system of values that make Chuck tick implanted those values in Beth after erasing hers. They did this while leaving her memory intact, which helps account for her surprise. Beth reflects on her new desire. Among other things, she judges, rightly, that it is utterly in line with her system of values. She also judges that she finally sees the light about morality—that it is a system designed for and by weaklings. Upon reflection, Beth has no reservations about her desire to kill George, is wholeheartedly

> behind it, and desires that her desire to kill George be effective. That desire is effective, and Beth succeeds in killing George.

In this story, the manipulators (the team of psychologists) not only succeed in getting Beth to kill her neighbor, George, but they also succeed in getting Beth to do this from a will that counts as free, according to Frankfurt's theory of freedom.[3] Upon hearing this story, however, many find it counterintuitive that Beth acts with free will. What cases like this show, according to Mele and others, is that freedom (and moral responsibility) is in some sense *history-bound*; for a person to act with free will, she must lack a history like Beth's—one in which the person has been manipulated in a certain way.[4] But since Beth satisfies Frankfurt's conditions on free will, the objector concludes, this case of manipulation is a counterexample to Frankfurt's account.

Before turning to how these types of cases are explored in Egan's story and in other works of science fiction, it is worth noting that cases of manipulation have also been used in the recently popular "manipulation argument" against compatibilism.[5] *Compatibilism* says that our being free is compatible with our being causally determined to act as we do by factors beyond our control.[6] According to the manipulation argument, agents who are manipulated lack freedom, yet there is no relevant difference between manipulated agents and ordinary agents who are causally determined by factors beyond their control, and thus compatibilism is false. If this argument is successful, not only compatibilist accounts that are structuralist (like Frankfurt's) would be undermined, but so too would *historicist* compatibilist accounts be undermined. In my defense of Frankfurt's account against the threat from manipulation below (in section 5), I will also suggest a response to this manipulation argument against compatibilism.

4. "Manipulated" Agents in Science Fiction

In Egan's "Reasons to Be Cheerful," Mark describes for us how, at the age of twelve, it is discovered that he has a potentially fatal brain tumor. Strangely enough, however, the tumor is causing pressure within Mark's brain (by blocking a ventricle), which is itself causing, in addition to some unpleasant symptoms like vomiting, "elevated levels of a substance called Leu-enkephalin—an endorphin, a neuropeptide which is bound to some of the same receptors as opiates like morphine and heroin" (Egan, 192). As a result, Mark is in a state of almost complete happiness, yet unless the tumor is removed, he will not survive.

Mark's parents arrange for him to undergo treatment to get rid of the tumor, and the procedure is successful. Even after seeing the test results showing that he is in the clear, however, Mark experiences nothing but a deep sadness after the removal of the tumor. "Everything I did," Mark tells us, "everything I imagined, was tainted with an overwhelming sense of dread and shame." Mark continues:

> The only image I could summon up for comparison was from a documentary about Auschwitz that I'd seen at school. It had opened with a long tracking shot, a newsreel camera advancing relentlessly towards the gates of the camp, and I'd watched that scene with my spirits sinking, already knowing full well what had happened inside. I wasn't delusional; I didn't believe for a moment that there was some source of unspeakable evil lurking behind every bright surface around me. But when I woke and saw the sky, I felt the kind of sick foreboding that would only have made sense if I'd been staring at the gates of Auschwitz. (197)

As it turns out, Mark's state of constant depression is the result of neurological damage, and nothing can be done for him for nearly two decades, until he is contacted by Dr. Durrani.

Durrani leaves a message for Mark explaining the treatment she has developed for neurological damage, and Mark recounts hearing the message:

> I listened as carefully as I could while Dr. Durrani explained her work with stroke patients. Tissue-cultured neural grafts were the current standard treatment, but she'd been injecting an elaborately tailored foam into the damaged region instead. The foam released growth factors that attracted axons and dendrites from surrounding neurons, and the polymer itself was designed to function as a network of electrochemical switches. Via microprocessors scattered throughout the foam, the initially amorphous network was programmed first to reproduce generically the actions of the lost neurons, then fine-tuned for compatibility with the individual recipient. (203)

The "network" that would act as a prosthetic for the lost neurons would allow Mark to experience pleasure again, without the overwhelming sense of dread and sadness that he had been experiencing for years. That network would itself be based on the neural networks of many cadavers with undamaged brains—four thousand of them, to be exact. When Mark asks about the nature of this network, Durrani offers a bit more explanation:

> We've used about 4,000 records from the database—all males in their twenties or thirties—and whenever someone has a neuron A wired to neuron B, and someone else has a neuron A wired to neuron C, you'll have connections to both B *and* C. So you'll start out with a network that in theory could be pared down to any one of the 4,000 individual versions used to construct it, but in fact you'll pare it down to your own unique version instead. (206)

Without an idea of what it will be like to experience this foreign network, Mark goes through with the procedure.

When he wakes, not only is Mark able to enjoy simple pleasures like the warmth of the bedsheets, but he also finds all of the people who gather around him stunningly beautiful. When he listens to music of any genre (and of any quality), he

rates it extremely highly (18 out of 20, or higher). He starts to worry that his tastes are not sufficiently discriminating, and he talks to Durrani, who tells him about the possibility of using software to allow Mark to "push" the network in one or another direction, with the result that he could directly control the degree of pleasure experienced in response to various inputs. The software is installed, and Mark begins using an imaginary "slider" to adjust his tastes and preferences. Most of his choices are motivated by pragmatics: rather than assigning a high degree of pleasure to junk food, he chooses "to crave nothing more toxic than fruit" (217); to put on some weight without risking a heart attack, he assigns a body type that is lean and wiry a 16 out of 20. And having been cured of the debilitating depression that had plagued him for years, Mark is able to secure a job at a local bookstore.

In working the bookstore counter, Mark interacts with many customers and begins to hope that he can meet someone that he could desire "more than all the rest" (Egan, 221). Mark recounts his predicament: "The 4,000 had all loved very different people, and the envelope that stretched between their farflung characteristics encompassed the species. That was never going to change, until I did something to break the symmetry myself" (221–222). And Mark does break the symmetry; after a week of adjusting his relevant systems such that he did not take any interest in anyone, Mark reverses these adjustments when a woman, Julia, enters the store. Mark recollects the experience:

> By the time she'd chosen two books and approached the counter, I was feeling half defiantly triumphant, half sick with shame. I'd struck a pure note with the network at last; what I felt at the sight of this woman rang true. And if everything I'd done to achieve it was calculated, artificial, bizarre and abhorrent, I'd had no other way. (222)

Mark asks Julie out to lunch, and the two begin seeing each other a couple times a week, at which point Mark starts reflecting more on what he is doing:

> Visions of Julia filled my head. I wanted to know what she was doing every second of the day; I wanted her to be happy, I wanted her to be safe. *Why?* Because I'd chosen her. But…why had I felt compelled to choose anyone? Because, in the end, the one thing that most of the donors must have had in common was the fact that they'd desired, and cared about, one person above all others. *Why?* That came down to evolution…my emotions had the same ancestry as everyone else's; what more could I ask? (223–224)

Although, at times, Mark feels like his choosing Julia was artificial, he also reminds himself that this is what choosing another person is like for everyone. "People make a decision," Mark muses, "half shaped by chance, to get to know someone; everything starts from there" (224).

Down the line, Mark tells Julia his entire story, including the details about the day they met. He apologizes, but Julia responds, "What are you sorry about? You chose me. I chose you. It could have been different for both of us. But it wasn't" (225).

The next day, however, upon further reflection, Julia changes her mind, as "she wasn't prepared to carry on a relationship with 4,000 dead men" (226). Mark is devastated, and although he can directly control his emotional response to the breakup, he lets things run their course.

At the end of the story, Mark receives a visit from his father, and, looking at his father, he takes the occasion to reflect on his condition:

> Watching him, I thought: he's there inside my head, and my mother too, and ten million ancestors, human, proto-human, remote beyond imagining. What difference did 4,000 more make? Everyone had to carve a life out of the same legacy: half universal, half particular; half sharpened by relentless natural selection, half softened by the freedom of chance. I'd just had to face the details a little more starkly. (Egan, 227)

Contrary to his earlier worries that his choosing of Julia was artificial, Mark comes to see that no person like us causes herself to exist without being influenced at all by legacy. We are all shaped, to some extent, by such factors as our evolutionary history, and so we must make choices from characters over which we lack total control.

Egan's story presents a thought-provoking case in which a person has direct control over one's own preferences, desires, and perhaps even values and character traits. Despite the differences between Mark's history and the history of typical persons like us, who come to form our preferences and values much differently, Mark comes to satisfy the conditions on freedom specified by accounts like Frankfurt's, so he counts as acting with free will when he chooses Julia.

Whereas Mark *has* (direct) control over aspects of his psychological profile, another type of agent, pervasive in science fiction, *lacks* control over its psychological profile, and yet this type of agent is, in a relevant sense, closely related to agents like Mark. I am talking, of course, about an artificial intelligence (AI), which is typically given its preferences, desires, etc. by its creators. The most interesting examples, for our purposes, are those in which an AI is made to satisfy conditions on freedom like those specified by accounts like Frankfurt's. Before turning to a response to the manipulation objection to such accounts of freedom, let us briefly consider Isaac Asimov's (1976) exploration of the freedom of an AI in his story, "The Bicentennial Man."

Artificial intelligences differ in degree of sophistication, but in many stories from science fiction they are created to be rather like human beings. Andrew, robot and main character of Asimov's story, longs to be more and more human. Along the way, having been a household robot for a number of years, Andrew wishes to become a free robot. The following is Asimov's description of an interchange between a Judge and Andrew during the hearing concerning his freedom:

> [The Judge] said, "Why do you want to be free, Andrew? In what way will this matter to you?"

> Andrew said, "Would you wish to be a slave, your honor?"
>
> "But you are not a slave. You are a perfectly good robot, a genius of a robot I am given to understand, capable of artistic expression that can be matched nowhere. What more can you do if you were free?"
>
> "Perhaps no more than I do now, your honor, but with greater joy. It has been said in this courtroom that only a human being can be free. It seems to me that only someone who wishes for freedom can be free. I wish for freedom."
>
> And it was that that cued the Judge. The crucial sentence in his decision was: "There is no right to deny freedom to any object with a mind advanced enough to grasp the concept and desire the state." (Asimov, 144)

This discussion of Andrew's "freedom" is political, not metaphysical; that is, what is at stake is Andrew's legal freedom, not whether he has free will (as if the latter could be decided in court). Nevertheless, Andrew's description of his desire for freedom strongly suggests that he is a sufficiently sophisticated AI that he satisfies Frankfurt's conditions on free will. Andrew has first-order desires to perform various actions, but he also desires that he bring about those actions in a particular way. This suggests that he is capable of reflection on his first-order desires and capable of desiring that one of his desires leads him to action, which just is what Frankfurt's account requires for free will. And even if one disagreed with Frankfurt about what was necessary for freedom, any structuralist account of freedom must admit that an intelligence constructed to be relevantly similar to ordinary human persons would not be any less free for having been programmed.

5. Responding to the Threat from Manipulation

Stories like Egan's "Reasons to Be Cheerful" and Asimov's "The Bicentennial Man" depict agents with abnormal histories—histories that may seem, at first glance, to be more like those of manipulated agents (like Mele's case of Beth) than like those of agents like you and me. But a common theme in these stories—one that is especially well-explored in Egan's story—is that these *prima facie* differences become less stark upon closer investigation. As we begin to reflect on the circumstances in which the characters of these stories find themselves, their experiences shed light on a certain aspect of the human condition, namely that we are bound to be shaped by legacy. And this leads to a response to the threat from manipulation to accounts of free will like Frankfurt's.

If persons like us ever begin to act with free will (which we must if free will is ever to get off the ground, so to speak), then our *first* of such actions must be from characters and preferences over which we did not have any control.[7] Presumably we perform such actions when we are young, though the exact age does not make a difference to the general point I am making. Of course, we typically go about performing actions that will shape our characters and preferences, taking ownership of some aspects of our initial psychological profile, rejecting others. At the beginning,

however, we are relevantly similar to agents like Mark and Andrew—we have been "given" a set of desires, dispositions, etc. that we did not select for ourselves. This is Mark's point when he says, "Everyone had to carve a life out of the same legacy: half universal, half particular; half sharpened by relentless natural selection, half softened by the freedom of chance" (Egan, 227). Given that we are agents with finite pasts, and given our lack of control over what we were like at the outset, the possibility of acting with free will depends on our being free at times when we acted from characters over which we had no control.

But notice that, once we see that we must allow for free will even in cases in which a person has no say over how she came to have the character she has, we must also admit that *manipulated* agents like Beth (from Mele's case) may act with free will. To be sure, Beth appears *less* free than a typical adult agent, but then again ordinary agents appear less free when they *start* to act with free will. A person's degree of freedom may increase over time as the person reflectively evaluates her starting point and begins shaping her psychological profile at later times (which sounds not unlike Mark's project). The point is that, so long as being manipulated does not prevent a person from satisfying structural conditions on freedom at the time of action, the manipulation does not undermine a person's freedom even if it diminishes her *degree* of freedom. In any case, we should not judge that manipulated agents like Beth lack freedom, *contra* the objection from manipulation.

It is worth pausing for a moment to say more about my invocation of *degrees* of freedom. The idea that freedom is a scalar notion (rather than simply a threshold notion, where it can simply be either "on" or "off") is widely accepted. It is common to think of freedom as a sort of control (e.g., control over one's behavior), and it is clear that control comes in degrees. For example, whereas a novice tennis player may exercise some degree of control over the placement of her serve, a more experienced tennis player may possess and exercise a much higher degree of the placement (and speed, trajectory, etc.) of her serve. In my view, because manipulated agents like Beth have less control over their own constitution (i.e., what they are like after being manipulated), they control their post-manipulation behavior to a lesser degree than do relevantly similar agents who have not been manipulated. Of course, even agents like Beth exercise *some* control over their conduct after being manipulated, and my view is that, insofar as such agent satisfy Frankfurt's conditions on freedom, they meet the threshold requirements for freedom (unlike non-persons). If a "manipulator" so altered a person's psychological profile that the person no longer satisfied even Frankfurt's structuralist conditions on freedom, such "manipulation" would indeed undermine the person's freedom. But the cases of manipulation and design that we have been discussing here are importantly different and seem to me only to mitigate agents' freedom.

This response to the manipulation objection to structuralist accounts of freedom may be extended into a response to the manipulation argument against compatibilism. Recall that, according to the manipulation argument, agents who are

manipulated lack freedom, yet there is no relevant difference between manipulated agents and ordinary agents who are causally determined by factors beyond their control, and thus compatibilism is false. Given the response to the threat from manipulation above, however, insofar as one is inclined to think that agents like us can possess free will, one should be prepared to grant that manipulated agents may be free, *contra* the first premise of the manipulation argument against compatibilism.[8]

In addition to its plausibility after reflection on the scenarios from Egan's and Asimov's stories, the response to the threat from manipulation that I have presented here is also consonant with Frankfurt's mature view. Consider the following passage from Frankfurt:

> A manipulator may succeed, through his interventions, in providing a person not merely with particular feelings and thoughts but with a new character. That person is then morally responsible for the choices and the conduct to which having this character leads. *We are inevitably fashioned and sustained, after all, by circumstances over which we have no control.* The causes to which we are subject may also change us radically, without thereby bringing it about that we are not morally responsible agents. It is irrelevant whether those causes are operating by virtue of the natural forces that shape our environment or whether they operate through the deliberate manipulative designs of other human agents. (Frankfurt 2002. 28, emphasis added)

Frankfurt's point, both here and in his earlier work, is that whether a person has free will (or is morally responsible) is not a matter of how one came to be as one is. After all, we are inevitably like Mark, or like Andrew, or even like Beth—fashioned by circumstances over which we have no control. Whether a person has free will, then, is only a matter of what they are like at the time of action, and the threat from manipulation fails to take seriously what all of us have in common with manipulated agents.

Before concluding, it is worth taking a moment to consider two potential objections to my argument. First, one might maintain that there is a crucial difference between cases like Mark's in which agents are aware of the origins of their psychological profiles, on the one hand, and cases of covert manipulation, on the other.[9] Perhaps acting with freedom of the will requires knowing the origin of one's psychological profile, the objection continues, and this requirement explains the difference between cases like Mark's and cases of covert manipulation. But if this objection were to succeed, it would prove too much, for ordinary agents (whom we take to act with freedom of the will) do not typically know the ways in which their psychological profiles were produced. To be sure, reflective agents recognize that they came into existence at an earlier point in time and that they began to make decisions and perform actions based on evaluative commitments that were influenced by a host of factors (including such things as parental influence, education, opportunities/lack of opportunities, etc.), but ordinary agents are not able to

determine the exact origins of the various psychological factors that lead them to perform specific actions. In fact, Mark's case is especially interesting because of its departure from the norm, so taking that case to exemplify necessary conditions on acting with free will would be to create very stringent requirements for so acting. Of course, one could take a strong stance here, accepting the implication that ordinary agents cannot act with free will, and this leads to the second objection we will consider.

A second potential objection to my argument sees the implications of my view for manipulated and designed agents (that they may act with freedom of the will) as a reason to doubt the notion of free will at all.[10] In particular, one might worry that if plausible accounts of freedom of the will (like Frankfurt's) must admit the freedom of manipulated agents, given their similarity to normal (i.e., non-manipulated) agents, perhaps we should conclude that even normal agents lack freedom rather than attributing freedom to both.[11] While this line of response may be tempting, and though some have endorsed it, I believe that it places unrealistic demands on freedom of the will and thus, ultimately, should not be accepted. According to the objection, to act with free will at a certain time *t*, an agent would have to have had control at some past time *t'* over what she was like at *t*. But in order for her to have been free at *t'* in exercising control over what she was like at *t*, she would have had to have had control at some even earlier time *t''* over what she was like at *t'*. And so on for eternity past. If finite agents like us (in particular, agents who began to exist) are ever to act with free will, however, it must be the case that we may do so despite not having total control over what we are like at the time of some of our actions. Requiring something like total control (or ultimate sourcehood) would be to add an impossible requirement on acting with free will, and it is my view that a proper assessment of our own limitations should motivate us to reject such stringent demands in favor of more modest conditions on acting with free will.[12]

6. Conclusion

I have argued that stories like Egan's "Reasons to Be Cheerful" provide the resources to respond to the main worry for structuralist accounts of free will (like Frankfurt's). Since those accounts say that having a certain psychological structure at the time of action is sufficient for having free will, an objection is that an agent may be manipulated into having such a psychological structure and would, because manipulated, appear unfree. I have argued, however, that we must allow for free will even in cases in which a person has no say over how she came to have the character she has, and admitting this should lead us to accept that manipulated agents may act with free will, *contra* the manipulation objection.

In addition to its defense of structuralist accounts of freedom, the argument of this paper has important implications for the freedom (and potentially *personhood*, insofar as it is connected with freedom) of artificial intelligences. Recall that, according to Frankfurt, "one essential difference between persons and other

creatures is to be found in the structure of a person's will" (Frankfurt 1971, 6). Frankfurt's account of free will is meant to characterize a distinctive feature of persons, and, as we have seen, it would be possible for a certain sort of AI (i.e., an AI capable of self-awareness and of desiring freedom) to satisfy Frankfurt's conditions. Having been created by another agent does not make a difference, on this view, to the freedom or personhood of the created agent; after all, we all began to exist at one point or another, and, like an AI, we did not have a say over what we were like when we created. Thus, despite the *artificiality* of an AI that satisfied Frankfurt's conditions, the AI would not be relevantly different from an ordinary agent with respect to its freedom and personhood.

Acknowledgments

I am very grateful to Eric Schwitzgebel for introducing me to "Reasons to Be Cheerful" by assigning it in his graduate seminar on AI Rights in the spring of 2016. Thanks to him and to the other participants in the seminar for feedback on the early stages of this project.

Works Cited

Asimov, Isaac. 1976. "The Bicentennial Man." In *The Bicentennial Man and Other Stories*, 135–173. Garden City, NY: Doubleday & Company, Inc.

Coons, Christian, and Michael Weber. 2014. *Manipulation: Theory and Practice*. New York: Oxford University Press.

Egan, Greg. 1998. "Reasons to Be Cheerful." In *Luminous*, 191–227. London: Millennium.

Feinberg, Joel. 1986. *The Moral Limits of the Criminal Law*, vol. 3: *Harm to Self*. New York: Oxford University Press.

Fischer, John Martin. 2011. "The Zygote Argument Remixed." *Analysis* 71: 267–272.

Fischer, John Martin. 2012. *Deep Control: Essays on Free Will and Value*. New York: Oxford University Press.

Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68: 5–20.

Frankfurt, Harry. 1988. *The Importance of What We Care About*. New York: Cambridge University Press.

Frankfurt, Harry. 2001. "Reply to John Martin Fischer." In *Contours of Agency: Essays on Themes from Harry Frankfurt*, edited by S. Buss and L. Overton, 27–31. Cambridge, MA: The MIT Press.

Jaworska, Agnieszka. 2007. "Caring and Internality." *Philosophy and Phenomenological Research* 74: 529–568.

McKenna, Michael. 2008. "A Hard-line Reply to Pereboom's Four-Case Manipulation Argument." *Philosophy and Phenomenological Research* 77: 142–159.

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

Mele, Alfred. 2016. "Moral Responsibility: Radical Reversals and Original Designs." *Journal of Ethics* 20: 69–82.

Pereboom, Derk. 2001. *Living without Free Will*. Cambridge: Cambridge University Press.

Pereboom, Derk. 2014. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.

Sartorio, Carolina. 2016. *Causation and Free Will*. New York: Oxford University Press.

Slote, Michael. 1980. "Understanding Free Will." *Journal of Philosophy* 77: 136–151.

Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies* 75: 5–24.

Todd, Patrick. 2013. "Manipulation." In *The International Encyclopedia of Ethics*, edited by H. LaFollette, 3139–3145. Blackwell Publishing Ltd.

Watson, Gary. 2004. *Agency and Answerability: Selected Essays*. New York: Oxford University Press.

Notes

[1] For another type of structuralist account, see Watson (2004).

[2] I focus here on Frankfurt's initial statement of his view and avoid sketching the details of subsequent modifications (none of which would be relevant here anyway). For Frankfurt's more fully developed position, see Frankfurt (1988), and for an excellent discussion and extension of the view, see Jaworska (2007).

[3] It is worth noting that such cases of "manipulation" are quite extreme and perhaps distant from the sort of thing we would typically call instances of *manipulation*. I am only interested in this more extreme type of manipulation here, but for further discussion of the nature of manipulation, see Todd (2013) and Coons and Weber (2014).

[4] For an early version of this objection to Frankfurt's theory, see Slote (1980). For more recent (and widely discussed) versions of the objection, see Fischer and Ravizza (1998) and Mele (2006). And note that while some "historicists" offer *negative* historical conditions like the one described in the body of the text ("for a person to act with free will, she must *lack* a history like Beth's"), some, including Fischer and Ravizza, proffer *positive* historical conditions, according to which an agent is free only if she *has* a certain sort of history.

[5] See especially Pereboom (2001; 2014) and Mele (2006).

[6] To be causally determined to act as we do by factors beyond our control is typically taken to involve (at the very least) the entailment of propositions describing what we do by propositions describing the intrinsic state of the world at some time in the distant past and propositions expressing the laws of nature.

[7] If a person had control, at some earlier time, over the character from which she now acts with free will, then she must have possessed free will at the earlier time, thus making the later instance of acting with free will not her *first*.

[8] This type of response to the manipulation argument has been called the "hard-line" reply, as it takes the hard line of granting that manipulated agents may free and even morally responsible for what they do. For influential developments of the hard-line response, see McKenna (2008) and Fischer (2011), and see Sartorio (2016) for an interesting error theory for the initial plausibility of the argument's first premise.

[9] Thanks to an anonymous reviewer for suggesting this point.

[10] Thanks to another anonymous reviewer for suggesting this objection.

[11] For a similar objection to the possibility of moral responsibility, see Strawson (1994).

[12] For similar responses to the objection to the possibility of moral responsibility, see Feinberg (1986, chapter 18) and Fischer (2012, chapter 10).