2013-05-01

# Responsibility, History, and Authenticity

Katcharov, Maxim

UNIVERSITY OF CALGARY

# Responsibility, History, and Authenticity

by

Maxim Katcharov

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF ARTS

DEPARTMENT OF PHILOSOPHY

CALGARY, ALBERTA

APRIL, 2013

**Abstract**

There is a debate in the literature on free will and responsibility regarding whether responsibility is historical. The focus is on what makes one's values, desires, beliefs, and other springs of action authentic or "truly one's own", and what effect, if any, past manipulation has on authenticity. In this thesis, I present and attempt to clarify fundamental concepts, and give a critical account of prominent theories and motivating cases. I propose constraints on theories of responsibility, including a generalized regress problem affecting "positive" historical theories, and argue for a characterization of springs on which springs of the same type are treated as identical. I conclude by presenting a sketch of a historical theory of responsibility which conforms to my proposed constraints.

**Acknowledgements**

Blessed be God in His Angels and in His Saints.

I am grateful for the help and guidance of my supervisor, Ish Haji, and for the Department of Philosophy at the University of Calgary.

I am especially thankful for the support of my family, and for Tess – her love and encouragement.

## Table of Contents

## Introduction

Actions arise under the influence of one's beliefs, desires, values, fears, and certain other mental states – in short, one's "springs of action". A person who, by acting, brings some event about might be responsible for that event. To what extent does such responsibility depend upon the history of springs, upon the way in which an action's springs themselves arose?

An ongoing debate in the study of free will and responsibility has led some philosophers to turn their attention to the role that history may play in a theory of responsibility. This debate has divided those philosophers into two main groups. *Historicists* have taken the position that relevant historical facts, particularly the facts of how one's springs of action were acquired at some past time, can render a person responsible or not responsible for an action. *Non-historicists* have taken the opposing position, roughly, that responsibility depends only on the facts of the present time. Within this debate, certain cases have been used to motivate the contrasting positions. Some cases, in which a person's values are surreptitiously modified, favour historicism. Others, in which a seemingly-responsible person has no history, favour non-historicism. Participants in the debate have used these cases to support novel criteria for moral responsibility.

This thesis is divided into three chapters. The first chapter provides an account of action and of closely related concepts, including basic causation, intentionality, consequences, responsibility, springs of action, authenticity, and history. The second chapter gives a critical account of various motivating cases and theories of moral responsibility, ranging from the "non-historical" theories proposed by non-historicists, to

the “historical” theories proposed by historicists. The third chapter concludes by establishing a framework of constraints on theories of responsibility in general and by presenting a sketch of a historical theory of responsibility.

## Chapter One: Background

In this chapter I will give an account of action and of closely-related concepts. The account will build up from events to actions and will then turn towards the causal antecedents of actions. The chapter will conclude by addressing the classification of certain relevant principles and theories.

As is the standard, I will give this account with reference to mental states, properties, events, and so on. Readers who are not comfortable with these terms may translate them to fit whichever neural or physical models they believe are most appropriate. Nothing substantial rests on differentiating between desires and the neural models of desires, to take one example.

The theory that I provide in this chapter and which serves as the basis for discussion in subsequent chapters is an *event-causal* theory. On an event-causal theory, events are modeled as being caused by prior events. I will not consider noncausal or agent-causal accounts of agency. On an agent-causal theory, some events are caused not by prior events, but directly by people (which is to say, by agents) as people or as substances. On a non-causal theory, some events are not caused at all, whether by prior events or by people. These agent-causal or non-causal views are therefore not open to the incompatibilist worries arising from determinism (to be discussed in the second chapter) that would pose a threat to event-causal theories. In contrast, actions arising under an event-causal theory seem to be restricted by various event-related natural laws. I take it that the relevant aspects of a plausible theory modeled on event-causation could readily be adapted to an agent-causal or non-causal theory of action. However, I will not undertake such a project. None of the philosophers whose contributions are discussed

below take up incompatibilist positions, and all propose or base their accounts on event-causal theories of action. I set agent-causal and event-causal theories aside.

### 1.1 Events and Causation

Imagine two ordinary, everyday events: Jill throws a rock, and Jeff's glass jar shatters. These two events are related: Jill's throw caused the shattering of the jar. The throw is the *cause*, the shattering is the *effect*.

There are coarser and finer ways to describe what has happened. At the coarsest level of description, we take the whole as one event: Jill breaks the jar. At a finer level, Jill makes a decision (1), which causes a signal to pass through her central nervous system (2), which causes bodily movement (3), which causes the rock to fly (4), and so on. This finer level of description is best-suited to the present discussion, as it neither omits nor obscures relevant detail.

In this four-event causal sequence, Jill's decision plays a specific role: it is the *direct cause* of the neural signaling, and thereby the *indirect cause* of the rock's flight. The story would be different, certainly not ordinary, but rather miraculous or paranormal, if Jill's decision had directly caused the rock to fly. Jill's decision can only indirectly cause the rock to fly. Indirect causation is not a secondary type of causation. To say that any event is the indirect cause of any other is just to say that there is a sequence of direct causes and effects joining the first to the second, and nothing more.

Causal relations are often much more complex than a linear sequence of four events. An event may have multiple direct causes and may bring about multiple direct effects. The number of indirect causes and indirect effects may be vast. Jill may indirectly bring it

about that Jeff's beetle escapes from his jar, that Jeff cuts himself while trying to catch it, that at a much later date Jeff casually chooses not to buy a jar of jam, and so on. Philosophers take pains to focus on simple linear causal sequences, but this is not to suggest that linear sequences are the norm. Such sequences are chosen for convenience, and the results seem to generalize.

Every change to a state, a property, or an attribute is marked by an event. When a jar that was unbroken becomes broken, this is marked by an event: the breaking of the jar. Some events seem to be instantaneous, occurring at an instant. Other events, such as certain processes, seem to occur over an interval of time. A rock may fly through the air, and this may take several seconds. When such an event stops occurring, there is an event that marks this termination of the event. The rock's flying through the air is terminated by the rock's colliding with the jar, which is an event. This distinction between events that occur over an interval of time and events that occur at an instant enables us to mark specific points in time when one process terminates and another begins.

An object may be conceived of or described (that is, subjectively "construed") in multiple non-interchangeable ways. The planet Venus can be seen as the morning star, or the evening star. A crumbling statue can be seen as partly destroyed, or partly intact. A lump of a white substance can be seen as a lump of sugar, or as a lump of poison. In the same way, an event may be subjectively construed in substantively different ways. Midnight can be seen as the end of the prior day, or the start of the next day. In such examples, the events may be identical, just as the morning star is identical to the evening star. However, the particular way in which the event is conceived by a person may play a pivotal role in our judgements.

Sometimes we speak as if particular objects were causes. We might say that the rock caused the jar to shatter. This is not strictly correct. The rock is not a cause: the rock is not an event. When we say that the rock caused the shattering, what we mean is that some event involving the rock has caused the shattering. The following is one plausible account. The rock's movement (an event) caused the rock's impact, and the rock's impact (an event) caused the shattering. This account does not require that objects serve as causes. Still, verbose descriptions of this sort, though precise, can be a burden. I will speak freely of rocks and other objects as if they were causes, with the expectation that a plausible causal story involving only events is apparent.

### 1.2 Intentions

A person can have various mental *attitudes* towards a particular event. A person can be aware that an event might occur, can want an event to occur, can be afraid that an event will occur, and can believe that it would be bad for an event to occur. A person can also believe that she can bring about an event. Some of these attitudes are epistemic, others are emotional responses, and others are reflective judgements. Attitudes are attitudes towards particular events: the key element of an attitude is that it has some particular event, subjectively construed in a particular way, as its *content*. Attitudes not pertaining to events might have other types of mental content: one might have attitudes towards a state of affairs or towards another person. For clarity I set those attitudes aside. In addition to the attitudes mentioned, a person can intend an event. Intentions, like the other attitudes, have some subjectively construed event as their content. However, they are a unique species of attitude because of the role that they play in action.

A person *intends* an event whenever she will operate in a way that brings about this intended event without the need for additional "deliberative thought".

This is a template definition of "intention": it captures a particular feature of the concept in question, but leaves open other important elements to further, potentially varied, definition. Its purpose is to capture the following point: once we form an intention, we tend to stop thinking about doing something, and instead start moving, in some appropriate sense of "moving". The definition leaves open precisely what counts as deliberative thought, but implies that deliberative thought of some variety must precede the formation of an intention.

On this account, *deliberative thought* is thought that serves to "form" or develop an intention. It is "deliberative" because it is in some sense responsive to circumstances, beliefs about the position of one's body, certain desires, and other attitudes. Deliberative thought may be prolonged and robust, but it may also be relatively short. At a minimum, however, deliberative thought is to be understood as being sufficiently robust so as to exclude involuntary reflex motions from counting as intended events. In particular, the sudden pulling-away of one's hand in response to extreme heat and pupil dilation in response to changes in lighting are not preceded by deliberative thought. In such cases, no intention is formed. A later section on deliberation will expand on what counts as deliberative thought.

The phrase "has an intention" should not be taken to necessarily imply that there must be some definite object that can be called an intention. A person does not necessarily "have an intention" in the sense that he may have a home or a coin in his pocket. Consider how a person might have a cheerful disposition, or might be having a bad day,

or might have an addiction to cocaine. In these senses of “having”, the person is in a certain state, but there is no definite object that might be called the cheerful disposition, the bad day, or the cocaine addiction. We may, however, still usefully speak of the development or formation of the disposition or the addiction.

A person who forms an intention has an intention at the moment that the intention is formed. Prior to this moment, the person did not have the intention. I am open to the possibility that after that moment, the person “had” the intention. In other words, it is sufficient for the account given that a person has an intention only at an instant. I discuss why I allow for this view in the next section. An analogy may be made to “having an idea”. A person who suddenly exclaims “I have an idea” means to say that he has just now formed the idea.

Intentions have a special role in the economy of action, distinct from other attitudes, because it is the content of the intention that determines which events were intended by a person. Although intentions have content, I do not I place them in the same category as desires, beliefs, values, and so on.

### 1.3 Resolutions

As defined, intentions are very much “immediate” or “proximal”. If movements do follow intention-formation, they follow almost immediately, and without the need for further deliberative thought. This formulation of intentions may seem to stand in contrast to a number of the typical uses of the term “intention”. It is normal for a person to say that he intends to visit his relatives at some relatively distant future date. Furthermore, such a visit does seem to require additional deliberative thought before it may be brought

about by the person. It is also normal to say that one intends to raise one's hand at a later time. I set such cases apart from cases of genuine intending, and refer to such attitudes as *resolutions*.

I contend that immediacy is a characteristic component of intention. Consider this personal experiment: resolve, now, to raise your hand tomorrow. In particular, resolve to raise your hand tomorrow as you might intend to raise your hand at this moment. Suppose that you do raise your hand tomorrow. This will happen in the following way. You will remember that you had resolved to raise your hand on the previous day, and then you will form the intention to raise it. Upon forming this intention, you will raise your hand. If you take raising your hand to be important, you may raise it very suddenly in response to suddenly remembering your resolution. If you take it to be unimportant, you may spend a few moments trying to recall exactly why you had resolved to raise your hand, and then you might raise it. In either case, the recollection of this resolution that you had made on the previous day would lead you, that morning, to form an intention to raise your hand. The resolution formed on the prior day acts in a similar manner to a desire or to a belief in leading you to raise your hand.

Now imagine that the following happens instead. Suppose that as you are walking about tomorrow morning, your hand suddenly shoots up (it shoots up exactly as your hand would shoot up now, if you had now formed the intention to raise your hand now). A lamp is knocked over by the motion, and you wonder why your hand is raised. Soon you remember: it is because you had formed a true intention, and not merely a resolution, to raise your hand on the previous day. This scenario is certainly a possibility: we might imagine a person with a bizarre neurological disorder that caused long delays between

forming intentions and moving as a consequence of forming those intentions. However, this is evidently not how we operate in all typical cases. This fact allows us to distinguish between intentions, which are characterized by immediacy (in normally-functioning human persons) and resolutions, which prompt one to form an intention in the manner that a desire might prompt one to form an intention.

It is certainly possible for one to resolve now to stop writing two hours later when the oven timer goes off. This would count as a resolution, not an intention. If one does respond to the oven timer, the response will necessarily have been preceded by what I have called "deliberative thought": thought that is responsive to, in this case, the oven timer going off, and which serves to produce (in a short span of time) an intention to abandon writing and check the oven. The person need not recall the precise history of how and for what reasons he formed this resolution. However, it is typically necessary to recall the content of the resolution, otherwise the person might find himself standing in the kitchen wondering what he had resolved to do, or perhaps even at his desk still writing and annoyed by the distracting noise produced by the oven timer, with only a vague feeling of having forgotten something.

I place resolutions in the same category as desires, beliefs, values, and so on. If, contrary to the points made in this section, one objects to the view presented and believes that resolutions truly are a type of intention, then one may freely refer to both intentions and resolutions as "intentions". But I treat the concepts as being distinct, and in the following sections I focus on intentions, and not on resolutions.

### 1.4 Forming Intentions

As defined, having an intention is a state that a person can be in: a person is in this state when she has a certain intention. A person enters this state of having an intention whenever an intention is formed. The formation of an intention marks the end of the process of deliberative thought (with respect to that intention), and the beginning of being in the state of intending.

People form intentions by deciding: a *decision* is an instance of active intention-formation. A decision is an event, and all instances of intention-formation are active (I will return to this point momentarily). Here the term "decision" is not used in the sense of coming to a conclusion or a reasoned judgement, but in strictly the preceding sense of forming an intention. A decision, on my view, is not the making of a choice between alternatives. A decision is, for lack of any better word, a sort of irrevocable commitment. Once a person decides, "that is it": the fact of the person having decided is permanently established.

I do not mean to imply that one might not at a later time form a contrasting decision. One may do so. The point that I mean to emphasize is that a decision is definite and has finality. In contrast, a preference may shift and change in intensity over time. A certain thought might arise repeatedly. But a decision is made once, at an instant, and without the possibility of future alteration or reproduction. A decision is not contingent on any outcome: it does not become any less a decision if the event intended in that decision fails to come about.

All instances of intention-formation are active, and all instances of intention-formation are decisions in the sense described above. This is because all intentions arise

through deliberative thought: thought that is in some sense responsive. The person is "involved" in all cases of intention-formation, and in cases where the person is not suitably involved, no intention is formed.

Consider a person who is driving down a highway and decides to take an exit, as she usually does. Her having decided entails that her intention arose as a result of deliberative thought. Clearly, her action was responsive. It is no coincidence that she formed the intention to take the exit: we may presume that she formed this intention in response to seeing the exit. Subsequently, she takes the exit, and she does so intentionally. As I have stated, there is a limit to what may be counted as deliberative thought. Consider a person who touches a hot stove, and reflexively pulls her hand back. This person does not form an intention, and does not pull her hand back intentionally.

The claim here is, ultimately, that all actions are intentional, though there may be unintentional passive events. (Actions will be discussed in a later section.)

### 1.5 Passive and Active

There is a difference between the things that we do and the things that happen to us. This difference is difficult to define, but easy to demonstrate. The following might merely happen to a person: she might inadvertently cough or sneeze, or she might breathe normally, or her heart may beat, or a strong gust of wind might cause her to stumble. These are all *passive*. None of these events are brought about by the person. Whether they are internal bodily processes, reflex motions, or movements produced by some external force, they are passive.

In contrast, the following might be done by a person: she might throw a rock, or she might run so as to make her heart beat faster, or she might intentionally cough, or she might suppress a cough, or she might subtract two numbers in her head. These are all *active*. Each of these is brought about by the person.

The language used in both passive and active cases is similar. In both cases we say "the person did". When we say "she coughed", it is not certain whether she coughed unintentionally, as due to a cold, or intentionally, as to draw attention to something. We must consider the circumstances, rather than the language used. This is not a trivial point: a number of non-trivial problems arise from insufficient attention to the question of whether something is done passively or actively.

We sometimes speak in a way that suggests that particular persons are causes. We might say something like "Jill broke the jar". When we say this, we do not necessarily mean that Jill, as an object, has caused the jar to break. We do not mean this for the same reason that we would not mean that the rock, as an object, has caused the jar to break. The rock is not an event, and neither is Jill.

However, when we say "Jill broke the jar", we also do not usually mean that Jill has played only a passive role in the causal sequence, perhaps by knocking the jar over after being blown off-balance by wind. When we say or imply that a person caused some event, what we typically mean is that she has played an active role, that she has exercised what we may call "direct control" over some event that played a role in the relevant causal sequence. Persons are not on this account causes, though it is sometimes convenient to speak as if they were.

## 1.6 Actions

An *action* is an *intentional* event, an event over which a person exercises *control* or agency.

Without a fuller account of control and agency, this may seem unsatisfying as a definition of action. Perhaps it is clear enough that people exercise control over their actions. But what are the unique qualities of agency and control? Such questions are too broad to even begin to address in sufficient detail. The point made here is roughly the following. If there is any event over which a person has control, any event that is suitably linked to intentionality, any event that is absent in all the aforementioned passive causal scenarios and is present in all the aforementioned active cases, then such an event is to be called an action. Furthermore, actions are intentional: they are appropriately related to what we recognize as intentions.

Although actions are intentional in a broad sense, some actions are not necessarily intended. Recall that a person who has formed an intention will operate in a way that brings about some intended event without the need for additional deliberative thought. The *intended* event is the event to be brought about. Raising one's arm, typing a sentence, and breaking a jar are all examples of intended events. Some intended events, such as bodily movements in particular, are often called actions. One common action that is intentional, but is *not* intended, is the mental action of actively forming an intention: the mental action of deciding. A person may resolve to intend (that is, resolve to decide and thereby act) at a later time, but a person does not intend to intend (that is, decide to decide) at a later time. Put plainly, when people act, they do not intend to form intentions, they intend to bring about events.

Some events can be described in multiple ways. With the same motion a person might either attempt to stretch her arm or attempt to hit someone. The person's intention strictly determines which action she performs: the intention determines which of these construals of the event counts as her action. It is possible that in addition to stretching her arm, she may unintentionally hit someone, but in such a case hitting someone would not count as her action. This is because it is not possible for a person to actively bring about an event without intending to bring about that event. In the case described, the person has no intention of hitting someone, therefore she is disqualified from performing such an action. We may, however, describe her as hitting someone, just as we may describe a person as coughing even when there is no intention to cough: the language used for speaking about such cases is ambiguous. Some of the confusion that may arise with respect to this point may be due to this issue of language. None of this demands that the person is an authority regarding her own intentions, since in some cases she may misunderstand her own intentions, or misremember.

### 1.7 Consequences

The *consequences* of an action are the events that are directly or indirectly caused by that particular action.

By virtue of exercising direct control over her action, a person exercises *indirect control* over all the consequences of that action. This includes all consequences that were intended, and all those that were not intended. As an alternative to saying that a person exercises indirect control, we may say that the person exercises control indirectly. Recall Jill, who threw the rock. By virtue of exercising direct control over her action, Jill

exercises indirect control over all the consequences of that action. She exercises indirect control over the neural signaling, her bodily movement, the shattering of the jar, and so on. She exercises indirect control over a vast number of other consequences that are not known to her, and impossible to enumerate here. These various outcomes of Jill's action were "up to her".

Much like indirect causation, indirect control is not a form of direct control, and it is not a second form of control. It is only a convenient way of saying that certain events were ultimately caused by an event over which the agent had direct control. As mentioned before, the story of Jill would be different, certainly not ordinary, but rather miraculous or paranormal, if Jill could exercise direct control over the rock's flight.

It is possible for a person to cause the actions of another person. Terry might ask Isaac to close a window, causing Isaac to close the window. Isaac's decision to close is a consequence of Terry's decision to ask. Terry exercises direct control over her own action and indirect control over Isaac's action, while Isaac exercises direct control over his own action. Actions are events; they are caused. Nothing prevents actions from having been caused by prior actions. In addition to exercising indirect control over the actions of others, a person may also exercise indirect control over her own future actions. Terry might write a note to herself, reminding herself to close her windows before leaving for a vacation.

An agent might or might not believe that her action will cause particular consequences. If an agent does not believe that (or does not know that, or is not cognisant that) an event will or may be caused by her actions, then, surely, the agent does not intend that event. Suppose that Jill knows nothing of neurophysiology at the time that she

decides to throw the rock. She does not know that particular cells in her body will transmit particular signals from her brain to her hand. So then Jill does not intend that her nerves fire, though neural firing is a consequence of her action. Suppose that Jill does not conceive of the possibility that her rock might bounce and hit Jeff. So then Jill does not intend to hit Jeff with a bouncing rock. Even awareness that a consequence may come about is not sufficient. Jill might believe that her heart rate will increase dramatically after she throws the rock. Increasing her own heart rate may be a foreseen consequence, but it would not count as Jill's intention.

### 1.8 Indirect Actions

I have mentioned intention-formation as an action that cannot be intended. Instances of intention-formation are decisions, decisions are *direct actions*, and (trivially) direct actions are instances of intention-formation. All three terms designate the same class of events.

On some broader accounts of action, certain bodily motions and other consequences resulting from decisions are considered actions. I refer to these as *indirect actions*. Not all consequences are indirect actions. I offer no criteria for distinguishing between those consequences that are and those that are not indirect actions. It is difficult to distinguish between indirect actions and consequences in general, and attempting to distinguish on the basis of bodily motion over which one seems to have control is problematic. When I now type, I have as much awareness of the movements of my fingers as I do of the movements of my muscles and the "movements" of my nerves. When I was first learning to type this was not the case: I would very deliberately move my fingers, with full

awareness of which keys I was moving them to. Yet as I type now, it seems to me that I have a form of control over my computer screen, perhaps just as it seems to me that I have control over the sounds that I hear myself making when I speak, and the motions I sense myself making when I move. Further difficulties arise for distinguishing between indirect actions and consequences when we consider persons with prosthetic limbs, who may believe that they have appropriate control of those limbs, even though those limbs are detachable mechanical devices.

Indirect actions (much like indirect effects and indirect control) are not a secondary type of action. Instead, they are a convenient way of speaking of various consequences of decisions, over which a person may believe that she has direct or indirect control.

### 1.9 Responsibility

Through the exercise of control, an agent may become responsible for certain relevant events. An agent is *directly responsible* for her authentic actions. An agent is *indirectly responsible* for the intended consequences of those actions. (A later section on the concept of authenticity addresses the question of which actions count as "authentic".)

On many accounts of responsibility, there are at most three conditions that must be met for a person to be responsible for an action or consequence. These are the epistemic condition, the control condition, and the authenticity condition. The epistemic condition requires that the person is in some sense aware of what the consequences of her action will be. If a doctor intends to cure a patient by administering a medicine, but the patient dies due to an obscure allergic reaction that the doctor was not aware of, then the doctor is not responsible for the death of the patient. According to the account I have presented,

the doctor is not responsible for killing the patient. Insofar as the doctor did not have the intention to kill the patient, the doctor *did not perform* the (indirect) action of killing the patient. Of course, the action that she did perform, which was administering medicine so as to cure the patient, caused the death of the patient. In this sense, the doctor did kill the patient. However, the doctor did not perform the action of killing the patient, so in another sense, the doctor did not kill the patient. The phrase "the doctor killed the patient" has two distinct meanings. The first is that the doctor performed the action of killing the patient. The second is that the doctor killed the patient passively or as an unintended consequence of some action. The death of the patient is a consequence of the doctor's action, but because the death of the patient is not an intended consequence, the doctor is not even indirectly responsible.

The control condition requires that an agent has appropriate control over the event in question. On the account I have presented, if an agent lacks appropriate control over an event, the agent does not perform an action. A person who lacks indirect control over an event does not have direct control over any antecedent cause of that event.

The authenticity condition requires, very roughly, that the action is "truly one's own", or that the person is the "ultimate originator" of the action, or that the action "belongs" to the person. Authenticity, we will see, has to do with the causal history of the causal antecedents, the springs, of action. Unlike the epistemic and control conditions, no relevant action is performed if the conditions are not met, an action that is inauthentic is still an action, but one for which the agent may not be responsible. The sections that follow will develop an account of the authenticity condition. It is enough for now to leave

authenticity as a "template": actions may be either authentic or inauthentic, and what counts as an authentic action is up to further definition.

In this thesis, I treat "responsibility", "responsibility for action", and "moral responsibility" as being effectively synonymous. I do believe that there is a distinction between responsibility for action and moral responsibility, but this distinction does not enter into the discussion in any relevant way. Part of the motivation for this distinction is that we may speak of a person being responsible for an action or event even if we do not know the moral status of the action or event (whether it is good, or bad, or forbidden, and so on). The following is a rough sketch of various conditions relevant to the concept of moral responsibility. One is *morally* responsible for an action or consequence if one is praiseworthy or blameworthy for that action or consequence. I set aside the concept of being praiseworthy. One is blameworthy for an action if one is responsible for that action and that action is morally forbidden. In this way, moral responsibility extends beyond responsibility, and beyond actions that have been performed to actions that have not been performed: one may be blameworthy for failing to act. While being morally responsible depends on an agent being responsible, nothing that I present will hinge on an agent being either praiseworthy or blameworthy in addition to being "just responsible".

### 1.10 Deliberation

Agents think before acting: they *deliberate*. Decisions and the intentional contents of those decisions arise through deliberation. During deliberation, agents exercise various capacities. They engage in basic reasoning on the basis of their beliefs. They make judgements on the basis of their values. They may have desires to act in opposition to

their values. Jill reasons that she can use the rock to break the jar, and realizes that Jeff will be upset if she breaks it. She judges that breaking the jar is not a nice thing to do, and she values being nice. She wants to seem tough in front of her new friends by breaking the jar, and acts on this desire despite her judgement.

Deliberation need not be robust. Jeff becomes aware that his beetle is escaping from the now-smashed jar, but does not want it to escape, so he decides to catch it. He makes this decision without thinking it through, and cuts his hand on the broken glass. Still, it is a decision, probably motivated by a strong desire not to let the beetle escape; it is nothing like a reflex. Many decisions are made after a comparatively short period of deliberation, though many others are made after a much longer period.

In contrast, when Jeff reflexively pulls his hand away after being cut, he does not deliberate, nor does he decide. One is not generally able to exercise direct control over reflexive bodily movements (though it is possible to anticipate such reflexes and to act so as to suppress them).

### 1.11 Springs

Desires, values, beliefs, resolutions, and other mental states affect the outcome of deliberation. Other mental states include expectations, predictions, hopes, emotions, fears, and opinions. Collectively, these are the agent's *springs of action*.

The content of the intention that is the outcome of deliberation will not necessarily match the content of every spring that played a role in deliberation. A person might fear being eaten by a lion, and yet after an extremely brief process of deliberation may form the intention to run away from the lion, which, of course, has a different content than the

aforementioned fear of being eaten by a lion. Some "non-attitudinal" springs might have a content that is not an event, and certain springs, such as a cheerful mood, may entirely lack content.

I leave open whether there must be some antecedent attitude whose content matches the content of the intention. However, I strongly suspect that there is such an attitude: it seems that the content of a resulting intention will match the content of some belief that played a role in deliberation, and that this belief will be a belief that the person has concerning what the person can do. A person might weigh a number of beliefs concerning what he can do in cases where he must make a choice between alternatives. If a person is trying to decide how he should escape from a lion, and believes that he is able to climb a tree and able to jump into a river, and in fact makes a decision, the content of the corresponding intention will match one of these beliefs concerning what he is able to actually do. In cases of simple decisions that do not involve a choice between alternatives, the content of the person's intention will likewise match some antecedent belief, such as the belief that one is able to turn one's vehicle to the right without driving into a ditch. However, I do not press this point.

Various moral qualities may be attributed a person in virtue of the way in which certain springs play a role in his actions. A person is diligent if he does not fail to act on his resolutions. A person is courageous (or foolhardy) if he fails to act on his fears. And so on.

For some people, the influence of certain springs is *irresistible*. A person who holds his breath will eventually be unable to continue holding his breath due to a reflexive response triggered by rising carbon dioxide levels in the bloodstream: this response will

manifest itself as an urgent desire to breathe. In such a case, the person would not be responsible for breathing, since he would lose direct control of his capacity to breathe. However, not all cases of action in which irresistible desires play a role lead to a person losing the capacity for direct control. A person who is addicted to cocaine might have an irresistible desire for cocaine, and consequently act on that desire. I take such a person to be at least potentially responsible for his actions, though not necessarily blameworthy.

Springs might arise in a number of ways. It would seem that springs generally arise in unremarkable and innocuous ways. However, springs may also arise through the exploitation of certain cognitive biases, or, conceivably, through the direct neural stimulation of a person's brain. In whatever way a spring is acquired, the agent may or may not realize that he is in the process of acquiring a spring at the time that he acquires that spring, and he may or may not later remember acquiring that spring. The agent may or may not have consented to acquiring some novel spring.

Springs might persist over various periods of time, and it may be easy or difficult to no longer have particular springs. Consider the following definition of unsheddability. "A pro-attitude is practically unsheddable for a person at a time if, given her psychological constitution at that time, ridding herself of that attitude is not a 'psychologically genuine option' under any but extraordinary circumstances" (Mele 1995, p. 172). Values are one type of "pro-attitude". Unsheddability applies equally well to other types of springs. We assume, as is the standard, that the various springs discussed are practically unsheddable, according to this definition.

### 1.12 Authenticity

Springs may be either *authentic*, or *inauthentic*. A carefully-considered judgement is a typical example of an authentic spring. A value implanted through the nefarious use of propaganda is a typical example of an inauthentic spring. Examples like these have motivated different characterizations of the authenticity criteria for springs. The following are two broad formulations of spring authenticity criteria. For reasons to be discussed in the following section, N-SA (below) is classified as a negative condition, while P-SA (below) is classified as a positive condition. Theorists typically endorse only one of these formulations, rejecting the other.

(N-SA) A spring is authentic only if it is not nefariously implanted.

(P-SA) A spring is authentic only if the agent has approved that spring.

I leave open what counts as "nefarious implantation" in N-SA. Typical examples include hypnosis, brainwashing, or the neurosurgical alteration of mental states for which the person in question is not responsible. On some accounts such examples are characterized as the *bypassing* of an agent's capacities of deliberative control. An agent who has had a spring nefariously implanted has been *manipulated* with respect to that spring.

Likewise I leave open what counts as approval in P-SA. Approval has been variously described as the consideration and mental "endorsement" of the spring, "identification" with the spring, or the claiming of "ownership" of the spring.

Actions may be brought about through the operation of many diverse springs. An action may be authentic or inauthentic (I leave open whether an action may be partly authentic). An action is considered an *authentic action* on the basis of the authenticity of

its springs. The following are three different rough formulations of action authenticity criteria:

AA1. An action is authentic if and only if *all* of its springs are authentic.

AA2. An action is authentic if and only if *any one* of its springs is authentic.

AA3. An action is authentic if and only if some significant proportion of its springs are authentic.

The motivation for the first formulation is that the influence of even one inauthentic spring is enough to make the action inauthentic. The motivation for the second formulation is that even one authentic spring is enough to make the action authentic. The third is the most likely candidate for a robust principle of action authenticity, but I will set aside any complete specification. It may be assumed that the various examples and cases discussed in this thesis will feature only actions that are entirely authentic or inauthentic, that is, actions brought about either entirely by authentic springs, or entirely by inauthentic springs.

### 1.13 Positive and Negative Theories

Any principle, theory, or thesis might be classified as positive or negative. A principle is a *positive principle* if it requires that an event has occurred (could occur, could have occurred). A principle is a *negative principle* if it requires that an event has not occurred (could not occur, could not have occurred).

These classifications are applicable to all principles. If a principle requires that a rock roll down a hill, then it is a positive principle. If a principle requires that a rock never has

the opportunity to roll down a hill, then it is a negative principle. These are trivial examples, but the classification applies to more prominent principles. Consider, as an illustration, the principle that moral "ought" implies "can". This principle states that a person is morally obligated to perform an act only if the person can perform that act. The principle that "ought" implies "can" is a positive principle, because its necessary condition is the possibility of an event occurring. (The "can" here does not express merely logical possibility.) A principle can fall under both classifications, so long as the occurrence of some event and the non-occurrence of some event are necessary conditions.

N-SA requires that no nefarious implantation of the spring has occurred, and is therefore a negative principle. P-SA requires that the agent has participated in establishing the spring or has approved of the spring, and is therefore a positive principle. (The *focus* of N-SA is on not having a spring nefariously implanted. The *focus* of P-SA is on the agent endorsing, identifying with, or otherwise "participating in establishing" the spring. These are two distinct explanations of our intuitions regarding cases of manipulation and defective springs. It is only incidental, though important, that one is a negative principle and the other positive.)

Using the terms "negative" and "positive" might suggest that the principles are negations of each other. But if a spring was not nefariously implanted, this does not entail that the agent has endorsed that spring. And if an agent has endorsed a spring, this does not entail that the spring was not nefariously implanted. If that were the case, then any dispute concerning the use of N-SA versus P-SA would be a dispute about whether a principle should be rendered "not x if y" (a spring is inauthentic if it is

implanted/unendorsed) or “x if not y” (a spring is authentic if it is unimplanted/endorsed). Clearly, such formulations are logically equivalent, and choosing one or the other is primarily a matter of personal preference.

Referring to P-SA as the positive principle also suggests that any similar positive historical principle is equivalent to P-SA. But this is not the case. P-SA is about something having actually happened: actual endorsement, identification, and so on. However, to take one example, McKenna sets out his positive principle PH (to be discussed later) so as to take into account only the possibility of an event occurring. This is not *the* positive principle, even though it is *a* positive principle.

Those points aside, I define “positive” and “negative” in the way that I do because it is not possible to “flip” events in the same way that properties can be “flipped”. The properties p and q, where q is not-p, are both properties. In contrast, an event which has occurred is an event, but an event not having occurred cannot be characterized as an event: it is always just a lack of an event, no matter how it is characterized.

Roughly, the point is that “has a certain type of history” (cf. McKenna forthcoming; Mele 1995, p. 172) cannot work as a characterization of positive principles, because both “was manipulated” and “was never manipulated” are types of history that a person can have. A principle that requires that an agent “was never manipulated” counts as a positive principle, on this characterization. But on an event-focused characterization of positive theories, while “manipulation” (“was manipulated”) is a type of event that could have occurred, “non-manipulation” (“was not manipulated”) is not in any sense a type of event that could have occurred.

Accounts of positive and negative principles that focus on properties and not events (or in general, on things like events that can be said to exist, or to have occurred) are not sufficiently precise. Consider the following candidate formulations:

> A positive historical principle specifies some particular kind of history that is required of an agent.
>
> A negative historical principle requires only that an agent satisfy the negative property of not having a certain kind of history.
>
> A negative historical principle requires that if an agent has a history, there is a property, *p*, which is such that the agent or an actional spring of that agent not have *p*.

Under these definitions, is the principle "a spring is authentic only if it lacks the property of being unendorsed" a negative or a positive principle? According to these definitions, this is a negative principle. Something should seem amiss, since this is, in effect, the positive principle. Yet nothing is amiss with the chosen principle. The property of "being unendorsed" is a perfectly serviceable property: the negation of any property is itself a property, and there are many properties that are natural negations of each other, such as "stationary" versus "mobile". Another example: as soon as manipulation occurs, a person loses the property of being free of manipulation, and gains the property of having been manipulated.

The principle "a spring is authentic only if it lacks the property of being unendorsed" is a positive principle. The definitions of positive and negative principles provided at the beginning of this section mark it as unambiguously positive, where other definitions fail

to do so. It is irrelevant that the principle contains the schematic phrase “lacks the property p”, since it defines authenticity as requiring endorsement, which is an event (to lack the property of being unendorsed is just to have the property of having been endorsed).

A convenient attribute present in the event-based classifications is that if the agent has no history, then the negative principle N-SA is satisfied, since if there is no history, then no manipulation has occurred.

### 1.14 Historical Theories

When we talk about past facts, those past facts are true now by virtue of certain circumstances or events having obtained in the past. The fact that Jill broke the jar at some past time is a historical fact, since it depends on it being true that in the past, a jar was broken by Jill. This is a relatively sterile form of historicism, perhaps “factual historicism”.

The type of historicism that is important to the various cases that will be discussed is more robust. It requires that facts of a present time depend both on other facts of that present time, and also on facts of a past time. Let us call these “robust historical facts”. The fact that in the past Jill broke the jar is not a robust historical fact, since it depends only on the past. It does not depend on whether the jar is present at the current time, or even on whether Jill is present at the current time. In contrast, the fact that Jill’s arm has a sunburn is a robust historical fact. It depends on there being a burn at the present time, and on the past fact that the very same burn was caused by the sun (rather than, say, a heat lamp).

A theory is a historical theory if it requires that some robust historical fact obtain. (The theories that will be presented will be either historical theories or non-historical theories that do not make reference to trivial historical facts.)

I turn briefly to certain historical qualities of springs. Suppose that you are working, and there is a glass of water on your desk. You briefly experience a desire to drink that water, but you decide to return to work instead of acting on that desire. You quickly forget about the water. Several minutes later, you remember your desire for a drink of water, but you realize the desire has now passed, that you "no longer have" this desire. You then glance at the glass, and again experience a desire to drink the water. This time, you decide to act on your desire.

The question is whether the desires mentioned are identical. If indeed they are identical, then you have felt the desire twice, remembered the desire once, chosen not to act on the desire once, and chosen to act on it once. If they are not identical, then it cannot be said that you have felt the desire before, but only that you have felt a similar desire.

These same points apply to other types of springs, such as values and beliefs. These points are important because we will want to ascribe certain historical facts to springs. We might want to say that a certain value was rejected by an agent, or that it was implanted, or that it was endorsed. But it would not make sense to say that a certain value was rejected if in fact we have good reason to believe that it was, in some sense, a "different" value that was rejected. I will return to these issues in the third chapter.

## 1.15 Review

An action is a discrete event, with preceding causes and consequent effects. In acting, agents intend to bring about certain consequences. Various springs of action (such as desires, beliefs, values and dispositions) factor into a process of deliberation. The decision that arises through this process of deliberation is a direct action.

There is disagreement as to how to formulate the authenticity conditions for responsibility. Some have proposed non-historical criteria, while others have proposed historical criteria. Of those who have proposed historical criteria, some support negative historical criteria, while others support positive criteria. We now turn to these views, and the cases that have been advanced to support these views.

## Chapter Two: Literature

In this chapter, I will give a critical account of various motivating cases and theories of moral responsibility. These will include Harry Frankfurt's historical theory, and the non-historical theories of Alfred Mele, of John Martin Fischer and Mark Ravizza, and of Michael McKenna. In response to Frankfurt's non-historical theory, I will set out Mele's case of manipulation, which elicits intuitions in favour of historicism (recall the cases of nefarious manipulation mentioned briefly in the above section on authenticity). In response to the historical theories, I will set out McKenna's case of an "instant agent", which elicits intuitions in favour of non-historicism.

Authors use varying terminology to discuss similar concepts. In many discussions, the concept of responsibility is intertwined and effectively equivalent to concepts of control and free will or free action. Such differences are marked where appropriate.

### 2.1 Determinism and Compatibilism

Determinism is the thesis that there is at any instant exactly one physically possible future (van Inwagen 1938, p. 3). We recognize that there can be many varying accounts of history: many varying accounts of what is true about the past. Two people can have conflicting beliefs about a certain day twenty years in the past. One person might believe that the sky was clear at some location, while the other might believe that it had rained all day. It is possible, as far as we can know, that either person is correct. Although we may acknowledge that both accounts of that day are a possibility, in one sense of "possibility", we tend to have a very clear idea that there is only one actual history. Determinism takes

this view with respect to the future: setting aside the various conceptual possibilities, there is only one future, just as there is only one past, and that future is determined.

On one popular account of determinism, this view of the future as being determined is grounded in a thesis which holds that a complete statement of all the laws of nature, together with a complete statement of the non-relational facts of the world at an instant in time, entails all future truths (Haji 2009 p. 18). As an illustration, consider the following variation of "War", a simple card game. The rules are as follows. Two players are each given a stack of cards. Each turn consists of both players revealing the top card in their stack. The card with the higher value is returned to the bottom of its player's stack; any card that is not returned is discarded. The player who first runs out of cards loses the game. Every outcome of this game is determined when the stacks of cards are assigned, even before the game begins. At every turn, the rules of the game, together with the positions of the cards in each stack, entail every future game-related fact.

Of course, the players are not strictly bound to follow the rules of the game. A player may cheat or abandon the game, or both players might agree to follow different rules. A typical proponent of determinism would contend that, in contrast to the non-binding rules of the game described above, the laws of nature are binding on normal persons, in some appropriate sense of "binding".

Some incompatibilists (that is, proponents of "incompatibilism") propose that alternative possibilities are required for responsibility: persons are responsible for having done something only if they could have done otherwise. Others also propose that responsibility requires that agents are the "ultimate sources" of their actions, but they argue that determinism eradicates alternatives or that it precludes the agent being the

ultimate source of her actions. So, they conclude, determinism is incompatible with responsibility.

Compatibilism is the thesis that is the negation of incompatibilism. It is the thesis that determinism is compatible both with free will and with responsibility.

### 2.2 Frankfurt

The debate regarding whether or not responsibility is in the relevant sense historical traces back to Harry Frankfurt's 1971 paper, "Freedom of the Will and the Concept of a Person". In that paper, Frankfurt develops an account of moral responsibility that appeals to the notion of second-order desires and argues that the distinction between persons and non-persons is that persons are able to form certain types of second-order desires.

Frankfurt motivates his view by considering cases of addiction. In particular, Frankfurt aims to provide a theory that will account for cases where a person, an addict, is overcome by some desire and seems not to be responsible. To avoid a complication, we assume that the addicts under consideration are not at fault for becoming addicts. Addicts, we will say, are persons who are periodically driven by intense desires to consume some drug. An unwilling addict is an addict who want these intense desires to fail to move him to action. This addict has a second-order desire that is in opposition to a first-order desire.

A first-order desire is a desire to perform (or refrain from performing) some action. When a first-order desire effectively moves (or will move) the person to perform the desired action, Frankfurt calls such a desire the person's "will". A second-order desire is a desire to have (or not have) a first-order desire. When a second-order desire is a desire

that some first-order desire be one's will (that is, that the first order desire would move one to action), Frankfurt calls such a desire the person's "volition". The will is the effective first-order desire, and the volition is the desire that some desire be effective.

Frankfurt refers to this coherence between volition and will as "identification". By having a certain volition, a person identifies with the first-order desire that the volition picks out. The first-order desire that the volition picks out is "truly the person's own". In cases where a person is unable to act from the first-order will that is "truly the person's own", the person fails to act freely, and consequently is not responsible.

> The unwilling addict identifies himself, however, through the formation of a second-order volition, with one rather than the other of his conflicting first-order desires. He makes one of them more truly his own and, in so doing, he withdraws himself from the other. It is in virtue of this identification and withdrawal, accomplished through the formation of a second-order volition, that the unwilling addict may meaningfully make the analytically puzzling statements that the force moving him to take the drug is a force other than his own, and that it is not of his own free will but rather against his will that this force moves him to take it. (Frankfurt 1971, p. 13)

In later work, Frankfurt expands on his theory and attempts to address a variety of objections. Detailed accounts of the various problems and responses can be found in Haji 2002 and McKenna 2011. I will not discuss the various responses and modifications here. Problems arise in cases where there are multiple volitions, and Frankfurt responds by attempting to clarify what it means for a person to identify with a desire. The basic form of Frankfurt's theory, however, remains the same.

Frankfurt's theory is non-historical. Of greatest relevance to our present discussion is Frankfurt's adamant response to the "manipulation cases". One prominent example of such a case will be set out in the section after the next. Manipulation cases aim to show

that various events in the past might have some authenticity-undermining effect on the person's springs of action. In response to such cases, Frankfurt maintains that:

> to the extent that a person identifies himself with the springs of his actions, he takes responsibility for those actions and acquires moral responsibility for them; moreover, the questions of how the actions and his identifications with their springs are caused is irrelevant to the questions of whether he performs the actions freely or is morally responsible for performing them (Frankfurt 1988, p. 54).

Frankfurt's theory is a positive theory. It requires the existence of entities with a definite ontological status: the first-order will and the second-order volition. For Frankfurt, a person is responsible only if a particular will and a particular volition exist.

### 2.3 The Drunk Driving Case

A first-shot against Frankfurt's non-historical theory typically involves a case called *Drunk Driving* (Mele 2009a, McKenna forthcoming). The case involves an agent who is directly responsible for deciding to become drunk to the point of losing control, and who subsequently becomes indirectly responsible for a consequence of that decision to drink. Consider the following brief exposition of the case by Mele:

> Van, a normal man, got drunk at a party and then tried to drive home. He was so drunk that he did not realize he was impaired. No one tricked Van into drinking alcohol, no one forced him to drink, he is knowledgeable about the effects of alcohol, and so on. Owing to his drunkenness, he drove into and killed a pedestrian he did not see. A plausible judgment about Van is that, other things being equal, he is morally responsible for killing the pedestrian. (Mele 2009a, p. 162)

Mele goes on to describe a similar man, Ike, who also drove drunk and killed a pedestrian. In contrast to Van, Ike was force-fed alcohol and placed into the driver's seat

of his car. Ike, we may reasonably conclude, is not responsible for killing a pedestrian. Van is responsible, but Ike is not. Mele contends that the relevant difference between the two agents is in how they came to be drunk, and that this difference is historical. I agree that there is something historical at play, but I do not believe that it is anything that would undermine, or even begin to undermine, Frankfurt's non-historical theory.

Consider the following historically-asymmetrical case, which I will call *Shot Down*. Jet and Mike are fighter pilots. Both are in flight over similar cities. Jet carelessly discharges his personal firearm, and this destroys his aircraft controls. Mike's airplane is hit by enemy fire, and this destroys his respective controls. Both pilots eject from their aircraft. The two scenes are now relevantly equivalent: as each pilot hangs in the air by his parachute, he sees his airplane crash into and destroy a priceless statue.

Directly before the airplanes crash, neither pilot is responsible for destroying a priceless statue. Consider the moment at which each airplane crashes into each statue. The "current time-slice" properties of both pilots and both airplanes are identical. However, at that moment only Jet becomes responsible for the destruction of a priceless statue. Mike does not become responsible, and for good reason: his airplane was shot down, and he was not the one who did the shooting. Should this case lead us to conclude that responsibility is historical?

To press this point further, consider the following case, which I will call *Ill Bill*. Jill carelessly gives Bill some bad grapes, which Bill consumes. Soon, Bill becomes seriously ill. Fifty years later, Jill has no memory of the event. She is, in all relevant ways, identical at that time to a person who had never made Bill ill. Yet she remains responsible for making Bill ill fifty years ago. What accounts for a potential asymmetric

judgement with respect to Jill? Here the explanation is clear: Jill remains responsible because she and not someone else carelessly made Bill ill fifty years ago.

No one of these three cases shows that responsibility is historical, in any robust sense of "historical". Jill, Jet, and Van are all responsible for the same reason. They are indirectly responsible for certain (potentially unforeseen) consequences of some prior decision. They are indirectly responsible by virtue of being directly responsible for certain actions that led to the outcomes in question. Had Jill not carelessly offered bad grapes, had Jet not carelessly discharged his firearm, and had Van not carelessly chosen to get drunk, then no one of these persons would be responsible.

In contrasting Drunk Driving with Shot Down, I am not drawing a comparison between the respective vehicles. Van and Jet were both careless, but while Jet lost control of his airplane due to his action, Van lost control of himself.

Mele seems, at least implicitly, to take for granted that Van drove actively: that certain motions resulted from intentional actions performed during driving. However, the case relies on Van having been entirely passive while driving. If either Van or Ike had had the appropriate level of control, then there would be no question of history. We would simply conclude that both Van and Ike were in control and thereby responsible for operating their vehicles without due care. There would be no asymmetrical judgement that could be used as leverage against a non-historical view of responsibility. The asymmetrical judgement relies on the passivity of the persons with respect to the events that they brought about.

The Drunk Driving case fails to show that responsibility is historical in any robust way. In contrast, the following case does pose a substantive obstacle to non-historicism.

## 2.4 The Ann-Beth Case

Mele's *Ann-Beth* case is intended to motivate a historical view. Ann and Beth begin as normal agents with different work-related values. Some third party nefariously alters Beth's work-related values to match Ann's work-related values. Subsequently, both agents make the same type of decision, and in doing so are influenced by values that differ only in how they arose. The following is Mele's popular account of the case:

> Ann is a free agent and an exceptionally industrious philosopher. She puts in twelve solid hours a day, seven days a week, and she enjoys almost every minute of it. Beth, an equally talented colleague, values many things above philosophy for reasons that she has refined and endorsed on the basis of careful critical reflection over many years. Beth identifies with and enjoys her own way of life, and she is confident that it has a breadth, depth, and richness that long days in the office would destroy. Their dean wants Beth to be like Ann. Normal modes of persuasion having failed, he decides to circumvent Beth's agency. Without the knowledge of either philosopher, he hires a team of psychologists to determine what makes Ann tick and a team of new-wave brainwashers to make Beth like Ann. The psychologists decide that Ann's peculiar hierarchy of values accounts for her productivity, and the brainwashers instill the same hierarchy in Beth while eradicating all competing values – via new-wave brainwashing, of course. Beth is now, in the relevant respect, a "psychological twin" of Ann. She is an industrious philosopher who thoroughly enjoys and highly values her philosophical work. Largely as a result of Beth's new hierarchy of values, whatever upshot Ann's critical reflection about her own values and priorities would have, the same is true of critical reflection by Beth. Her critical reflection, like Ann's, fully supports her new style of life.
>
> Naturally, Beth is surprised by the change in her. What, she wonders, accounts for her remarkable zest for philosophy? Why is her philosophical work now so much more enjoyable? Why are her social activities now so much less satisfying and rewarding than her work? Beth's hypothesis is that she simply has grown tired of her previous mode of life, that her life had become stale without her recognizing it, and that she finally has come fully to appreciate the value of philosophical work. When she carefully reflects on her values, Beth finds that they fully support a life dedicated to philosophical work, and she wholeheartedly embraces such a life and the collection of values that supports it.
>
> Ann, by hypothesis, freely does her philosophical work; but what about Beth? In important respects, she is a clone of Ann – and by design, not accident. Her own considered values were erased and replaced in the

> brainwashing process. Beth did not consent to the process. Nor was she even aware of it; she had no opportunity to resist. By instilling new values in Beth and eliminating old ones, the brainwashers gave her life a new direction, one that clashes with the considered principles and values she had before she was manipulated.

The case has been presented in full. Beth has been manipulated, and has acquired a new configuration of work-related values in a troubling way. This manipulation subverts not only her immediate springs of action, but also her capacity for critical reflection. Although she is capable of critical reflection, the relevant values that play a role in reflection have been constructed in such a way that Beth wholeheartedly embraces all of her new values. Subsequently, she decides to do philosophical work. Mele continues:

> Beth's autonomy was violated. And it is difficult not to see her now, in light of all this, as heteronomous – and unfree – to a significant extent in an important sphere of her life. If that perception is correct, then given the psychological similarities between the two agents, the difference in their current status regarding freedom would seem to lie in how they *came* to have certain of their psychological features, hence in something *external* to their present psychological constitutions. That is, the crucial difference is *historical*; free agency is in some way history-bound. (Mele 2006, p. 164-6; originally in Mele 1995, p. 145-6)

Mele's intention is to make a strong case for historicism, and he does so by motivating the intuition that Beth is not responsible. The following three points seem to lead us to think that Beth is not responsible. The first is the purposeful involvement of another agent. The dean decides to exercise indirect control over Beth's action, intending to bring it about that Beth would perform the action. Due to the dean's calculated interference, Beth does as the dean had intended. Second, Beth had identified with and endorsed her former values. She considered and explicitly rejected the dean's suggested values. Her former values were then replaced by those new values.

These two points do not explain our intuitions. Suppose that the dean was able to change Beth's values through normal modes of persuasion, such as through discussion. It would then be true that another agent had become involved and had caused Beth to have new values. It would also be true that certain values, with which Beth had strongly identified, were replaced by new values that she had initially rejected. However, neither of these two points suggests that Beth's subsequent actions would have been defective.

The third and crucial point is that the manipulation bypasses normal modes of acquisition. Beth's mental configuration seems to have been established in an abnormal way. This is not merely a matter of the values having been implanted directly, or of the values having been constructed. We might imagine a case where Beth had commissioned the brainwashers herself, and in such a case, we would not consider Beth's subsequent actions to be defective. What seems to matter is that the new values were implanted, first, without Beth's consent, and second, without her awareness. That is, normal modes of acquisition involve (at least) Beth being aware of the acquisition and consenting to it. The direct implantation, in this case, fails to meet these criteria.

The case is structured so as to undermine two potential responses. First, Beth is changed in only a limited sphere of her life. This makes it difficult to argue that Beth is not responsible due to having become a new person through a radical and complete change in personality. Second, the case gives a plausible account of how Beth fails to realize that she has been manipulated. This makes it difficult to argue that the case is implausible because an agent could not be manipulated without becoming aware.

Finally, we should be clear about what Beth is and is not responsible for. Beth is a talented philosopher. When she writes a philosophical paper, she is in the relevant sense

responsible for the work that she produces. If she writes the paper, she would be the author, and not anyone else. The source of the ideas presented in the paper would undoubtedly be Beth. What is at issue is Beth's choosing to work on the paper. Likewise, if an athlete were brainwashed into competing, she would be credited with her performance, but she would not be responsible for entering the competition. The judgement that Beth is not responsible only applies to her choice to do work.

Frankfurt responds to manipulation cases in the following way:

> Briefly, it seems to me that if someone does something because he wants to do it, and if he has no reservations about that desire but is wholeheartedly behind it, then–so far as his moral responsibility is concerned–it really does not matter how he got to be that way. One further amendment must be added to this: the person's desires and attitudes have to be relatively well integrated into his general psychic condition. Otherwise they are not genuinely his, but are merely disruptive intruders on his true nature. As long as their interrelations imply that they are unequivocally attributable to him as his desires and attitudes, it makes no difference–so far as evaluating his moral responsibility is concerned–how he came to have them.
>
> 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. We are the sorts of persons we are; and it is what we are, rather than the history of our development, that counts. The fact that someone is a pig warrants treating him like a pig, unless there is reason to believe that in some important way he is a pig against his will and is not acting as he would really prefer. (Frankfurt 2002, 27-8)

Evidently, Frankfurt remains committed to his non-historical position despite the intuitive pull of manipulation cases.

## 2.5 Mele

In *Autonomous Agents,* Alfred Mele supplies the following spring-authenticity condition. Although Mele notes that the condition (below) may not cover all cases of inauthenticity, he does believe that it is sufficient "for present purposes" – purposes that at least include the discussion of the Ann-Beth case. To say that an agent is compelled* to possess a pro-attitude is, roughly, just to say that the agent has an inauthentic spring.

> If an agent *S* comes to possess a pro-attitude *P* in a way that bypasses *S*'s (perhaps relatively modest) capacities for control over his mental life; and the bypassing issues in *S*'s being practically unable to shed *P*; and the bypassing was not itself arranged (or performed) by *S*; and *S* neither presently possesses nor earlier possessed pro-attitudes that would support his identifying with *P*, with the exception of pro-attitudes that are themselves practically unsheddable products of unsolicited bypassing; then *S* is compelled* to possess *P*. (Mele 1995, p. 172)

Mele's condition is series of four conjoined clauses, or sub-conditions. The second clause states that the agent is not excused from being responsible in cases where she intentionally brought about her own manipulation. The third clause states that the agent is excused in cases where, roughly, she would have been able to abandon or "shed" her implanted springs, but had failed to do so.

The fourth and final clause states that an agent is not excused in cases where she presently possesses or earlier possessed attitudes that would support identification with the implanted springs. The principle allows that implanted attitudes may be authentic by virtue of the agent's possessing (though not exercising) some collection of attitudes through which the agent might identify with the implanted values. Put plainly, if the agent has some collection of values that could support newly-implanted values, then identification with those novel implanted values is effectively implicit. It seems that

implicit identification or endorsement may be problematic, but I will not press this point. Of greater importance is the matter of possessing such values at some past time. I will return to this point in the third chapter.

What remains is the initial clause: a spring is not authentic if the agent came to possess that spring in a way that bypassed his capacities for control over his mental life. This is, I think, the crucial point of Mele's negative principle. Mele explains capacities for control and bypassing in the following way:

> In some cases of brainwashing, hypnosis, and the like, agents come to possess pro-attitudes in ways that *bypass* their (perhaps relatively modest) capacities for control over their mental lives. In ideally self-controlled agents, these capacities are considerable. Such agents are capable of modifying the strengths of their desires in the service of their practical, evaluative judgments, of bringing their emotions into line with relevant judgments, and of mastering motivation that threatens (sometimes via the biasing of practical or theoretical reasoning) to produce or sustain beliefs in ways that would violate their principles for belief acquisition and retention. They are capable, moreover, of rationally assessing and revising their values and principles, of identifying with values of theirs on the basis of informed, critical reflection, and of intentionally fostering new values and pro-attitudes in themselves in accordance with their considered evaluative judgments. Presumably, most readers of this book have each of these capacities in some measure.

Mele continues:

> All such capacities are bypassed in cases of pro-attitude engineering of the sort at issue. In such cases, new pro-attitudes are not generated via an exercise or an activation of agents' capacities for control over their mental lives; rather, they are generated despite the agents' capacities for this. (Mele 1995, p. 166-167, footnote omitted)

Mele lists various mechanisms by which an agent could exercise control over desires, emotions, motivations, and beliefs. His claim is that in cases of manipulation, at the moment that novel attitudes are nefariously implanted, these various capacities for control are not effective (see the last sentence of the above quotation). This seems

reasonable. Mele is not saying that various springs are authentic only if exercise or activation occurs, since this would make his theory a positive theory. However, the emphasis on exercise or activation seems out of place in a negative theory since these are events, and this seems to leave manipulation underspecified.

We may presume that not every acquisition of springs in which the agent fails to exercise or activate these capacities counts as a case of manipulation. Mele is willing to leave it open that an instantaneous agent is morally responsible for her first decision. As an example, he describes Athena, who "magically comes into existence with a wealth of beliefs, desires, and values in place" (p. 1995, p. 172). "Other things being equal", Mele writes, "there is no bar to Athena's autonomously possessing her pro-attitudes". Mele reports that he has doubts about whether Athena is a possible being (p. 172). But he proposes that if she is possible, it looks as though a being may autonomously possess values she has over a stretch of time that begins at some moment, even if she had no history at all as an agent before that moment.

In this case, it seems that a certain relevant event occurred: the creation of Athena's values in such a way that there was a failure to exercise or activate relevant capacities. This failure was due to the fact that no such capacities existed at that time, since Athena was only at that moment coming into being. It is not clear on what basis cases like this do not count as cases of manipulation. This is not intended to be a serious objection against Mele. The point here is simply that a specification of manipulation itself would be helpful (though it is perhaps understandable that it would be difficult to specify in isolation from a specification of what counts as the proper exercise or activation of various faculties for acquiring springs).

### 2.6 Fischer and Ravizza

In *Responsibility and Control*, John Martin Fischer and Mark Ravizza develop a historical account of responsibility. Broadly, it says that persons are morally responsible for actions that causally issue from reasons-responsive mechanisms or processes for which they have taken responsibility. The account has two parts. The first part, *reasons-responsiveness*, is a specification of the control condition: it is a specification of the sort of freedom that responsibility requires. This aspect of their account will not be relevant to our discussion. The second part, *taking responsibility*, advances a historical view. On Fischer and Ravizza's account, people are able to take responsibility for certain mechanisms of action:

> The process by which an agent takes responsibility for the springs of his action makes them *his own* in an important sense (Fischer & Ravizza, 1998, p. 210).

Taking responsibility is a matter of meeting three conditions. The following is a rough sketch of these conditions. The first condition requires that the person considers herself to be an agent. That is, she views herself as being able to bring about certain events in the world. This first condition calls to mind the behaviour of very young children, who may move, but who do not yet appreciate that their movements produce various effects in the world.

The second condition requires that an agent see herself as a valid target of moral appraisal. This condition is complex, and I will not examine it in detail. One interesting upshot of this second condition is that certain people are excused from being responsible, but at the cost of no longer being considered moral agents:

> At a minimum, we expect our attitudes toward others to have some purchase. The fact that we might further blame someone who fails to react appropriately toward himself indicates that *we* believe the person is an appropriate candidate for the reactive attitudes. If, however, a person resolutely shows no moral response or *appreciation* of the moral force of the attitudes we take toward him, then eventually we must concede that he is not an appropriate partner in the conversation: he has not taken responsibility for himself. In this case, we stop resenting him as a person, and begin treating him as we would a distasteful object or a dangerous (or annoying) animal. (Fischer & Ravizza 1998, p. 213)

Recall that Frankfurt takes a similar approach in dealing with persons who do not form second-order volitions. Such persons are, according to Frankfurt, not really persons at all. Fischer and Ravizza apply this strategy in dealing with cases in which a person is simply unable to form appropriate conceptions of himself as a moral agent. They offer the following analogy, which invokes the image of a person out at sea:

> The basic idea is that an individual who really does not see himself as an agent and a fair target for the reactive attitudes cannot be deemed genuinely active and morally responsible. In *not seeing himself* in a certain way, he *fails to be* a morally responsible agent. Lacking the required view of himself, he *is* essentially passive, buffeted by forces that assail him. (Fischer & Ravizza 1998, p. 221)

The analogy here may seem confused. In the relevant sense, even a normal, responsible person is buffeted by forces that assail him. In deciding, he forms intentions, and consequently a "force" moves his body. This "force" (now departing from the analogy) effectively traces back to the exercise of the person's own capacity for direct control. The case of the person who fails to take responsibility seems to not differ in any relevant way. What forces buffet this self-deluded person, if not his own direct agentive control?

In connection with this condition, Fischer and Ravizza mention a case of a serial killer, Michael Ross, who describes himself as not being in control of his thoughts and

actions. Ross seems to attribute these actions to a "monster" that is dwelling within him. A person who is faced with the fact of having done a terrible wrong might respond by distancing himself from his own actions, and attributing them to some other force – a monster, an intense desire, fate, deterministic atomic processes, and so on. A person's conception of himself may be deluded, and he may believe that his movements must arise in a way outside of his own direct control. However, this does not seem to be sufficient reason to conclude that the person's movements are in fact produced by some mechanism that lies outside of his own direct control. Fischer and Ravizza's third condition, however, seems to address these concerns:

> Finally, the cluster of beliefs specified by the first two conditions must be based, in an appropriate way, on the individual's evidence. (Fischer & Ravizza 1998, p. 238)

I take this to mean that if a person is not justified in taking such a view of himself, then that person might very well be responsible. If there is not some responsibility-subverting process operating within the agent, then the agent is likely not justified in believing that there is such a process, and is therefore responsible. If this is the track that Fischer and Ravizza take, then a clearer approach may be to set the agent's self-conception aside, and stipulate that what matters is whether the person really *is* an agent, and really *is* an appropriate target of moral appraisal. I put this issue aside.

When an agent does take responsibility, he takes responsibility for a mechanism of action. This is effectively equivalent to endorsing certain springs of action. If certain springs arise through some mechanism, such as nefarious manipulation, for which the

person does not take responsibility, then those springs are inauthentic, and the person is not responsible for actions produced by those springs.

This does not, however, preclude the possibility that the agent has taken responsibility for the mechanism as a result of manipulation. Fischer and Ravizza concede this point:

> On our approach, an agent need not be held morally responsible for acting on certain moderately reasons-responsive mechanisms that have been "implanted": these mechanisms are not the agent's own. But it is conceivable that a different sort of manipulation takes place, in which the agent's taking responsibility itself is somehow electronically implanted. That is, it is conceivable that the individual's view of himself as an agent and an apt candidate for the reactive attitudes be electronically implanted. Does our account of taking responsibility (and moral responsibility) imply that such an agent must be considered morally responsible?
>
> Earlier, we specified the third condition on taking responsibility as follows: the agent's view of himself must be based on his evidence in an appropriate way. Obviously this is abstract and schematic. This condition is intended (in part) to imply that an individual who has been electronically induced to have the relevant view of himself (and thus satisfy the first two conditions on taking responsibility) has *not* formed his view of himself in the appropriate way. But the relevant notion of appropriateness must remain unanalyzed. (Fischer & Ravizza 1998, p. 236)

Fischer and Ravizza go on to state that although they are not offering a knockdown argument for their theory being compatible with determinism, they have nevertheless made such a claim highly attractive. This unanalyzed notion of appropriateness, however, is a weak point of Fischer and Ravizza's theory. Manipulation cases were developed in part to undermine a particular mechanism: a process involving Frankfurt's second-order volitions. In such cases, the manipulated person's endorsement of various springs is itself inauthentic. In other words, the mechanism by which an agent takes responsibility is itself implanted. When manipulated Beth takes responsibility for her newly-implanted springs, it seems clear that she does so purely in virtue of the fact that she has been manipulated. Had she not been manipulated, she would not have endorsed those springs.

Imagine that Beth did realize that she was manipulated, and understood that her springs were inauthentic. Furthermore, imagine that, owing to the influence of these inauthentic springs, she did not consider the fact that she acquired them in the way that she did to be problematic. She might now believe that it is appropriate to manipulate persons in a way that causes them to work on philosophy. In such a case, Beth would have an appropriate epistemic view of how her springs came about and of what had happened to her, and yet we may reasonably presume that she has not formed the view in an appropriate way, on Fischer and Ravizza's account of appropriateness.

Without a specification of this notion of appropriateness, Fischer and Ravizza's account of taking responsibility seems open to objections that arise even within fairly typical manipulation cases.

### 2.7 The Suzie Instant Case

Recall the Ann-Beth case, which was presented in favour of a historical view. The following is a summary. Ann values philosophy and is eager to do philosophy, while Beth values other things above philosophy and does not work so diligently as a professor. A third party subjects Beth to nefarious manipulation: Beth's former work-related values are replaced by Ann's values, and Beth is oblivious to what has occurred. This change causes Beth, upon critical reflection, to endorse this new style of life and to work as diligently as Ann does. Ann seems to be responsible for choosing to work on philosophy, but Beth seems to not be responsible. Since Beth is in all relevant ways the present-time duplicate of Ann, the differential judgement must arise from historical considerations: in

particular, from something to do with the fact that Beth was nefariously manipulated at an earlier time, while Ann was not.

McKenna considers the Ann-Beth case to be compelling, though not definitive. He believes that a similarly-compelling case can be presented in favour of non-historicism: he describes the case of *Suzie Instant*, who spontaneously comes into being as the duplicate of Ann (and post-manipulation Beth), and then acts.

> Suzie Instant is created by a god at an instant. She is created to be a psychologically healthy woman indistinguishable from any other normally functioning thirty-year-old person whom any of us might encounter. To get this result, she is given a huge set of beliefs according to which she has lived a normal human life for thirty years. For instance, she believes (falsely) that she had a twelfth birthday and that her daddy bought her a pony. Furthermore, Suzie has some range of unsheddable values. She also has a set of false beliefs about how she came to acquire those values. She thinks that she acquired them through a process of sustained effort over the years leading up to what she thinks is her thirtieth. She takes pride in this fact and believes that she is responsible for this process and that she engaged in it freely. (On this point, clearly she is mistaken.) She is, also, a richly self-controlled person who is able to resist the inclination to act with weakness of will. When she acts, the desires issuing in her actions are the ones she wants to act upon, and when she does, she is sensitive to a wide range of reasons for action. Hence, Suzie satisfies an impressive set of features of the sort that, when she acts, varying theorists would regard those features as adequate for satisfying all of the nonhistorical conditions highlighted in their respective accounts of free and morally responsible agency.
>
> Suppose that Suzie is presented with the option to do one of two things, A or B. One option, B, involves a violation of a value that is unsheddable for her. The other option, A, involves acting from one of her unsheddable values. Suzie A-s, acting as her unsheddable value counsels, but in doing so, she could have done otherwise – that is, she could have B-ed. It is difficult to see how a causal history that zeroed in on Suzie Instant all in an instant renders her unfree in a way that she would not be if instead some causal history or other unfolded over the course of thirty years. Note that when Suzie A-ed from her unsheddable value, being able to do otherwise, she was not compelled to do so. Her A-ing was *nothing like* acting upon an irresistible desire. It would be natural to say that she A-ed freely – in at least some non-question begging, restricted sense of freely, say freely*. (McKenna forthcoming)

Suzie's values were created when Suzie was created, through an act of a divine being. Suzie has no prior values that are overwritten. In the Ann-Beth case, the dean had intended that Beth perform an action. To avoid an irrelevant complication, we may suppose that Suzie's creator did not intend that Suzie perform any specific action.

McKenna proposes that we should be inclined to take Suzie Instant to be responsible. She is, after all, in no way different from Ann (who is responsible), save that her history "came compressed in a momentary package". Here, McKenna argues for there being no clear relevant distinction between a person who has come into existence at an instant, and a person with an arbitrarily short history. Thus Suzie is a free agent, but only by virtue of her non-historical properties, since, McKenna argues, she has no history. Suzie having the same non-historical properties as Beth gives us reason, according to McKenna, to treat them as either both responsible, or both not responsible.

One problem with McKenna's case is that it is so far apart from reality that our intuitions may be misled in some way. At the cost of introducing several confounding points, a more plausible variation of the case is as follows. Imagine that a person suffers an unexpected brain injury that so affects his personality and in such a way that there is an overwhelming intuition that he is no longer the same person. This person essentially comes into being at the moment of the injury. When this new person thereafter acts, intuitions (or at least my intuitions) lean towards holding this new person responsible for his actions. The same would apply to an agent like Suzie Instant (or Mele's Athena, mentioned above), since she is a new person in the relevantly same way. It does not matter whether it is ultimately correct to say that the person is not the same, so long as we can have sensible intuitions about the case; as long as we can accept that the person is not

the same for the sake of the discussion, where in the case of Suzie Instant it may be difficult to do so.

The Ann-Beth and the Suzie instant cases elicit conflicting intuitions. Reflection on the Ann-Beth case elicits the intuition in many that manipulated Beth is not responsible for her actions that issue from those springs that had been nefariously implanted. She is not responsible because, in some sense, these springs are "not her own". The case suggests that the way in which one acquires one's relevant springs of action plays a role in appraisals of responsibility. In contrast, although Suzie Instant seems responsible for her germane action, she has no past. Hence, this case is intended to pull in the other direction: that appraisals of responsibility are not dependent on history.

McKenna does not commit to a non-historical account, and later argues for the plausibility of a historical account. I turn to this historical account in the next section.

### 2.8 McKenna

McKenna's paper, "A Modest Historical Theory of Moral Responsibility" (forthcoming), contributes to the compatibilist debate concerning historicism and non-historicism. McKenna prefaces his main points with several introductory remarks.

The first introduces direct and derivative responsibility. As McKenna defines the terms, an act is directly free if it "issues from a direct exercise of the free will ability", and indirectly free if its status as free "traces back in a suitable way" to a directly free action. The case that McKenna invokes in favour of derivative responsibility is a typical variation of the Drunk Driving case, which we have discussed in a prior section.

The second remark introduces Mele's idea of "thinly valuing x". One thinly values x if two conditions are met: one must have a "positive motivational attitude" towards x, and believe x to be good. In other words, of the things that one is motivated to pursue, values are those things that one considers good. If a person thinly values x, then x is a thin value for that person. The third remark makes reference to Mele's concept of unsheddable values. A value is unsheddable over a period of time if, over that period of time, one can do nothing to reduce the intensity of that value, or the extent to which that value will affect one's action. All of the values that McKenna's paper is concerned with are unsheddable. The fourth remark introduces the notion of being free to do something, or being able to do otherwise. The point is that the values that lead agents like Ann, Beth, and Suzie to action must not be so strong that they rob the agent of free will. They must not be overpowering.

All values at issue are not overpowering, unsheddable, and are at least thin values. I will use the term "value" without specifying that the value is a thin value that is unsheddable and not overpowering.

### *2.8.1 NH and PH*

It seems that a manipulated agent like Beth is not responsible for those actions that are the result of manipulation. McKenna discusses two principles that may account for this intuition. He refers to these principles as the negative historical constraint NH, and the positive historical constraint PH. McKenna ultimately rejects NH in favour of PH.

We will examine the way in which McKenna has formulated these principles, and then consider his reasoning for rejecting one in favour of the other. The primary aim will

be to clarify the two principles by separating action authenticity conditions from spring authenticity conditions. A secondary aim will be to clarify the principles by revising McKenna's chosen wording, again without altering the intent of the principles. McKenna's formulations are relatively long, and the details are complicated. McKenna formulates the two principles as follows:

> NH: An agent A-s freely and is morally responsible for doing so only if, with respect to the causal springs of her A-ing, she does not have a history that includes the acquisition of any unsheddable values through means that bypassed her ability to critically acquire, assess and sustain them.
>
> PH: An agent performs a directly free act and is directly morally responsible for it only if any unsheddable values playing a role in the production of her action arose from a history whereby she was afforded the opportunity to critically assess, endorse, and sustain them from abilities that she possessed, and so none were acquired through means that bypassed those abilities. (McKenna forthcoming)

Authenticity conditions for actions are distinct from authenticity conditions for springs. The above principles specify both conditions within each principle, but this is unnecessary. The action authenticity condition is effectively identical for both principles. McKenna's action authenticity condition is as follows:

> An agent performs an act freely and is morally responsible for performing that act only if all the causal springs of that act were authentic.

This condition is effectively the AA3 action authenticity condition specified in the first chapter. Under this condition, an agent is absolved of responsibility if any spring is inauthentic, even if that inauthentic spring plays only a minor role in deliberation and action. This view is problematic, since it is easy to imagine a case in which some minor belief or desire has been nefariously implanted and plays a causal role in the action, but

in which the person would have acted in the same way even in the absence of the implanted spring due to other pre-existing springs that were influential and authentic. McKenna's focus, though, seems not to be on the authenticity conditions for actions, and we will set these concerns aside.

The focus is on the authenticity conditions for springs. McKenna's NH and PH principles differ primarily in their spring-authenticity criteria. The principles also differ in the abilities that are listed as playing a role. Both principles list the agent's abilities to assess and sustain relevant values, but NH lists the ability to acquire values while PH lists the ability to endorse values. We will examine NH first, and then PH. The spring authenticity condition provided in NH is as follows:

> A value is authentic only if the agent does not have a history that includes the acquisition of that value through means that bypassed her ability to critically acquire, assess and sustain that value.

This formulation is problematic. McKenna emphasizes the agent as having a history, but it is the history of the agent's spring that is at issue; the agent is parenthetical. Where the principle states "only if the agent does not have a history that includes the acquisition of that value", we may state "only if it was not acquired", since this is equivalent.

A value cannot be acquired in a way that bypasses the agent's abilities to acquire that value. Much depends on the word "critically", since otherwise we are left with the impossible scenario in which a value is acquired, but acquired in a way that bypassed all means of acquiring the value. McKenna does not elaborate on the difference between acquisition and critical acquisition, but the difference seems to be that the agent plays no active role in acquiring the value. The difficulty of this point may be due to the same

difficulty of language that makes it unclear whether, when we say that a person coughed, we mean that she coughed actively or passively. In this scenario, McKenna seems to be suggesting that though the agent acquires the value passively, the agent does not acquire the value actively, which is to say that her ability to "critically acquire" is bypassed.

We will not focus on which particular abilities McKenna believes play a role in nefarious implantation. In the context of McKenna's principles, let us define "supporting" as follows. To *support* a value is for an agent to critically acquire (or endorse), assess, and sustain that value from abilities that she possessed. The following phrasing of the spring authenticity condition for NH is effectively equivalent to the original:

> A value is authentic only if it was not acquired through means that bypassed the agent's ability to support that value.

The concerns listed above have to do with the way in which McKenna has chosen to define the template term "nefariously implanted". McKenna, in line with Mele, takes nefarious implantation to consist in acquiring values through means that bypass certain abilities.

Examining McKenna's PH lends credibility to the above revised formulation of NH. The final clause of his PH principle, which originally reads "none were acquired through means that bypassed those abilities", is effectively identical to the above revised formulation of NH. McKenna states that "PH builds upon NH because it includes the negative proviso featured centrally in NH – lacking a history whereby the pertinent values were acquired by coercive means". The revised NH is precisely what McKenna

has in mind when he mentions the "negative proviso featured centrally in NH", which we have called the spring authenticity condition for NH.

We turn now to McKenna's positive spring authenticity condition. The spring authenticity condition for PH is as follows:

> A value is authentic only if the value arose from a history whereby the agent was afforded the opportunity to critically assess, endorse, and sustain that value from abilities that she possessed, and so it was not acquired through means that bypassed those abilities.

As before, the phrasing "only if the value arose from a history whereby the agent was afforded" may be replaced with the equivalent "only if the agent was afforded", and for the same reasons. When McKenna uses the term "afforded", he does not mean that the agent was explicitly "granted the opportunity", as if by another agent. He means that the agent merely had an opportunity. As before, these concerns have to do with the formulation rather than the content of the principle. The following two concerns present difficulties that are more serious than difficulties of interpretation.

PH differs from NH in two puzzling ways. First, there is potentially a reference to two distinct levels of abilities. In NH, the agent has an ability to support values, and this ability is bypassed. In PH, the agent has the opportunity to support values "from abilities that she possessed". I take this to mean that the agent possessed the abilities to assess, endorse, and sustain (and furthermore had the opportunity to exercise these abilities), and not that the agent exercised those abilities on the basis of some set of secondary abilities. This leads to the following re-formulation of the principle.

A value is authentic only if the agent had the opportunity to support that value (and so it was not acquired through means that bypassed her ability to do so).

Earlier I mentioned that the last clause of PH was similar to the whole of NH. The second puzzling difference is that McKenna's PH seems to include NH as a conjunct. If this is so, then PH is a stricter principle than NH, since any value that is inauthentic according to NH will also be inauthentic according to PH. One possible interpretation is that McKenna has included this cause only parenthetically, and that he takes this clause to be only an implication of the principle. But if this is the case, it is best to omit this clause entirely. Because of these concerns, I will set aside the extra clause and consider only the central positive component of McKenna's PH. The central spring-authenticity components of McKenna's principles are as follows, with the action-authenticity condition and NH repeated for convenience:

An agent performs an act freely and is morally responsible for performing that act only if all the causal springs of that act were authentic.

(NH) A value is authentic only if it was not acquired through means that bypassed the agent's ability to support that value.

(PH) A value is authentic only if the agent had the opportunity to support that value.

It may be helpful to compare these formulations of the principles to McKenna's original formulations, which can be found at the start of this section. Recall that in this context, for an agent to support a value is for an agent to critically acquire (or endorse), assess,

and sustain that value from abilities that she possessed, and that we have omitted the final clause of McKenna's original PH.

McKenna's PH differs in a substantive way from the positive principle given earlier as P-SA. It does not require that an agent actually support some relevant spring, such as a value, but requires only that an agent had an opportunity to do so. We will see why McKenna introduces this opportunity clause in the next section.

McKenna remarks that in order to avoid a regress in any positive historical thesis, assessing, endorsing, and sustaining must not themselves be construed as actions, or as events that depend on earlier free actions. McKenna states that he makes no claim regarding what event it must depend on, but then also makes the claim that it depends on having the opportunity to critically assess the value. In a later footnote, he states roughly that it depends on actually exercising one's ability. In the conclusion of his paper, he seems open to an account on which the event that must occur might merely be having an opportunity to exercise one's rational capacities.

### *2.8.2 Beth Passive*

McKenna objects to NH on the basis of what it implies with respect to the following case, involving an agent named *Beth Passive*:

> Beth Passive simply lived out her life passively allowing her interests to be whatever they happened to be. She inherited them almost exclusively by aping her parents, and then later some high school and college chums. It never once crossed her mind to consider whether there were other modes of life that are more rewarding, such as one that involves a singular commitment to the life of the mind, much like the kind of life Ann has fashioned for herself. Beth Passive, it turns out, is no more than a value-sponge. Had it been that her parents were more like Ann, Beth Passive would likely have sponged up those values instead. It might well be that, like Beth Active, Beth Passive is morally

> responsible for the character she has acquired, but her responsibility consists largely in hapless omissions.
>
> NH yields the result that, when Beth Passive is manipulated into being like Ann, she does not act freely nor is she morally responsible. This is just what the historical theorist wants. But the problem is that it is hard to see *why* NH should apply to Beth Passive in a way that it is not at all hard to see why NH should apply to Beth Active. Beth Passive blindly stumbled into her moral personality. If we are to regard as freedom-and-responsibility undermining the intervention into Beth Passive's psychic life whereby very different unsheddable values are covertly forced upon her, ones upon which she subsequently acts, it is not because Beth Passive was robbed of a moral personality that she came to possess under her own steam. (McKenna forthcoming)

Unlike the original Beth, Beth Passive does not refine her values on the basis of critical reflection over many years. She has, instead, resorted to a much less demanding method of acquiring values: she simply adopts the values of the people around her without discrimination. Then, as described above, Suzie Instant's values are replaced by a new set of values, without her consent or participation.

A negative principle like NH holds that it is objectionable to bypass an agent's normal modes of acquiring values. McKenna agrees that Beth is wronged, and he does see the implantation of new values into Beth Passive as responsibility-undermining, but not because they bypass Beth's (indiscriminate) mode of acquiring values. According to McKenna, responsibility is undermined because Beth Passive loses the opportunity to endorse her own values.

### *2.8.3 Loss of Abilities*

McKenna claims that PH renders Beth not responsible for her actions. For this to be the case, it must be true that Beth has lost the opportunity to support her new values. However, while Beth gains new values, she does not lose her abilities to support her new

values. The Ann-Beth case stipulates that Beth explicitly embraces, which is to say supports, her new values:

> When she carefully reflects on her values, Beth finds that they fully support a life dedicated to philosophical work, and she wholeheartedly embraces such a life and the collection of values that supports it. (Mele 2006, p. 164-6)

Beth does meet McKenna's positive requirement of having the opportunity to endorse her new values, because she in fact did endorse her new values. We must turn to some other aspect of the PH principle to see why Beth would not be responsible. Has manipulation robbed Beth of "abilities that she possessed"? Perhaps, though Beth supports her new values, she does so out of abilities that she does not, in some sense, possess. The answer hinges on how we are to interpret "abilities". If by "abilities" McKenna is referring to Beth's capacity to reason, her agency, and so on, then it seems that Beth retained her abilities. After all, the case stipulates that the only qualities lost by Beth were her former work-related values.

So perhaps by "abilities", McKenna has in mind Beth's work-related values or abilities nontrivially associated with these values. In what sense is Beth no longer in possession of her work-related values? The only plausibly available sense is the usual sense – she has been manipulated, and that the values are therefore "not her own". Presumably the point is that unless an agent has authentic values which might inform her capacity to support (or endorse) values, she never has what can be counted as an opportunity. Her newly-implanted values cannot serve this role, since they are not authentic. Her former values cannot serve this role because they have been obliterated.

## Chapter Three: A Framework for a Historical Theory

In the first sections of this chapter I argue for certain constraints on positive historical principles. On my view, these constraints render theories that rest in whole on a positive principle highly problematic, or even untenable.

I then argue for a view of springs and authenticity that does not rely on particular token springs, but instead treats springs in a "general" way. Roughly, the view is that springs are more appropriately viewed as being more akin to types than to tokens. As an illustration, consider how a person might be said to like a story. A story can be told in many ways, and on many different occasions, but the person would recognize that it is the same story being told each time. The person might like the story "in general", but have no opinion of any particular telling of the story. I argue for a similar treatment of springs.

I conclude the chapter by presenting a basic historical theory that is in line with the views that I have argued in favour of.

### 3.1 The Regress Constraint on Positive Principles

Recall that a positive authenticity principle requires that some event had occurred (or could occur, or could have occurred). Certain positive authenticity principles suffer from a regress problem. Consider P-SA:

(P-SA) A spring is authentic only if the agent has approved that spring.

Now consider the following principle:

Any spring approved through an inauthentic action is itself inauthentic.

This is a reasonable principle that proposes that manipulation is “contagious”: that any spring or process that is established on the basis of other inauthentic springs or processes is itself inauthentic.

Recall that an inauthentic action is an action performed on the basis of inauthentic springs. If approval is taken to be an action, then it must be performed on the basis of prior authentic springs. It follows from this that no spring can be authentic unless it is preceded by an authentic spring. This generates an infinite regress: there can exist no authentic spring that is not itself preceded by an infinite sequence of antecedent authentic springs. Therefore, for any positive principle, the required event that establishes the authenticity of a spring cannot itself necessarily be an action, or in general any event for which an agent is responsible.

This problem of there being an infinite regress that results from taking the relevant event to be an action is relatively well understood. McKenna is cautious to avoid such a regress involving one’s actions and springs of action, and he stipulates that endorsement cannot be construed as an action.

### 3.2 A Generalized Constraint on Positive Principles

The relevant event (call it endorsement) in a positive principle cannot be an action due to the regress problem described in the previous section. Nor can this be any event for which the agent is responsible, for the same reason. The goal, then, for any positive theory is to specify what sort of event could count as endorsement. There is, however, a limit to what sort of events can reasonably be counted as endorsement. Suppose that this non-action is any specific event that arises, in some typical way, on the basis of the

agent's values (or other springs). If this action does depend on such values, then it must require that such values be authentic. Otherwise, it would be possible to construct a counterintuitive manipulation scenario in which inauthentic springs bring about this (non-actional) endorsement event, which then produced authentic values. So this event, if it depends on any values at all, must depend on authentic values. But if the values that underlie this event must themselves be authentic, the question arises just as in the case of actional endorsement: how can any effective value be rendered authentic, except through some prior event of endorsement? This is the same sort of regress problem encountered in the previous section. The upshot is that endorsement cannot strictly depend on *any* antecedent values, whether inauthentic or authentic. In the same way, endorsement cannot strictly depend on *any* antecedent desires, beliefs, fears, or other springs.

To be free of an infinite regress, a positive principle must specify endorsement as a non-actional event for which an agent is not responsible, and which must not arise through the influence of any antecedent springs, whether those springs are inauthentic or authentic. This leaves very little room for a plausible view of endorsement. The problem with positive accounts, then, is that they require some pre-existing and non-agentive faculty of endorsement that is immune to manipulation. It is difficult to imagine what this faculty could be, given that people are, at least apparently, not immune to manipulation.

Negative theories do not suffer from this problem, because they require only that manipulation has not occurred.

### 3.3 The Nature of Springs

Manipulation cases depend greatly on what it means for two springs to be identical. Consider the following case, which I will call the *Fran-Beth* case. New-wave brainwashing works in the following way. To begin, brainwashers enter the novel hierarchy of values into a computer console. The console is wired up to an oversized ray gun, which is concealed behind a curtain. The subject is led to a spot marked "x", and the ray gun is activated. When the brainwashing rays hit the subject, her certain former values are obliterated, and the novel values are implanted in their place.

Fran is the psychological twin of Ann, and the physical twin of Beth. She is a diligent philosopher, a colleague of Ann and Beth, and the sort of person who would never consent to any form of manipulation. Deep in thought, Fran wanders through a room containing the concealed ray gun. She notices that someone has scrawled an "x" underfoot, and stops to scuff it out with her shoe. A novice brainwasher mistakes Fran for Beth, and activates the ray gun. Fran's former work-related values are obliterated and replaced by the new values. The new values, however, precisely match Fran's old values. Fran feels mildly dazed, and she decides to return to her office to work on philosophy.

It would seem that Fran has been manipulated. There is a brief moment between having her values obliterated and having ostensibly novel values implanted in which Fran is certainly deprived of her values. For the purpose of generating a contradiction, suppose that the springs that Fran had before being blasted with the ray gun are *not* identical with the springs she has after being blasted. If they are not identical, then Fran no longer meets either the positive or the negative spring authenticity condition. Since the values are (by assumption) novel, Fran has not considered and endorsed these values, since she acts

immediately. Furthermore, Fran lacks even the mere opportunity to endorse (or evaluate, or assess) these novel values, again because she acts immediately. It is also true that the (by assumption) novel values were acquired through nefarious means that bypassed Fran's normal modes of acquiring values. Fran is, in these two crucial ways, the historical duplicate of manipulated Beth. However, Fran *does* seem to be responsible for deciding to return to her office and work.

My proposed solution is to reject the assumption made in the preceding paragraph. The springs that Fran had before being brainwashed must be identical with the springs she had after. Springs are not to be conceived of as concrete particulars, as if residing in the agent's brain or subconsciously in the agent's mind. When Fran in some way supports a particular value, she does not support some particular configuration of neurons. When a person endorses a value at one time, considers it at another time, and acts on it at a third time, the value is the same value throughout.

One might insist that the values are different, and consequently, that the values are inauthentic, and that Fran is not responsible. I do not believe that this is a tenable position.

Consider first the following case, which I will call *Immediate Cure*. This case is intended to be similar in form to Mele's original Ann-Beth case. Ann and Beth both undergo a psychological evaluation in order to determine what their values are. Beth is then manipulated. Her former work-related values are obliterated, and values that are of the same type as Ann's values are implanted. However, immediately after manipulation, the new-wave brainwashers are apprehended. The brainwashers are forced to re-implant values that are of exactly the same type as Beth's original, pre-manipulated values, which

they have “on file” due to their earlier psychological evaluation of Beth. The procedure is, thankfully, reversed. However, if one insists that values are different, then there is no “cure” for Beth. Like Fran, she has been manipulated (twice), and her relevant subsequent actions are inauthentic, since they stem from inauthentic springs.

Furthermore, consider the following case, which I will call *Nazi Propaganda*. Imagine a loyal Nazi, with the standard set of Nazi values, which he freely developed over a period of time. This Nazi freely chooses to watch a film series that he thinks that he will enjoy. This film series is actually extremely potent Nazi propaganda. The Nazi does not realize that the films are propaganda, and does not set out to be manipulated. The films, however, turn out to be extremely effective, and for nefarious reasons: they exploit a vast number of human cognitive biases, provide false but seemingly authoritative information, and so on. The first films are a general, nihilistic attack on pertinent values, and they cause this Nazi to shed these values. However, the films eventually change course, and in the same nefarious way they implant the standard set of Nazi values. Immediately after manipulation is complete, the loyal Nazi performs some action on the basis of these (by assumption) novel values. Perhaps he salutes and proceeds to sing and march about his apartment for several hours, as he usually does, which greatly annoys his neighbours. If these new values are not the same, then they are inauthentic, in which case this Nazi is no longer responsible for his actions – by virtue of freely choosing to watch a series of films that were strictly aligned to his own authentic values.

Might this issue be addressed by maintaining that post-manipulation values do differ from pre-manipulation values, and by stipulating that “novel” values are authentic if and

only if “prior” values of the same type were authentic? Perhaps, but it is puzzling why this approach would be appealing. If by virtue of (say) endorsing one token value of a certain type a person endorses all token values of that type, then does this differ in any substantive way from simply endorsing the “type”? It is not different, but only more complex.

Springs, then, are more like types; they are not tokens. This has several interesting implications. Among these implications is the idea that it would be more appropriate to say that springs may become authentic or inauthentic.

### 3.4 Dependence on Last Instance of Acquisition

Consider the following scenario, which I will call *Cult*. Kurt was once a member of a peculiar cult. The sole tenet of this cult was that the titles of books placed vertically on a shelf should be read by tilting one’s head to the left, not the right. At that earlier time, Kurt had voluntarily endorsed and adopted as a value this singular tenet of the cult. Cultists, Kurt among them, would invade libraries and turn the vast majority of shelved books upside-down. Eventually, Kurt was captured by librarians. Through the legitimate use of reason and civilized discourse, the librarians convinced Kurt to abandon his former book-inverting value. Kurt became a nuisance to cultists, and they hired brainwashers to re-instill Kurt’s former value.

Kurt returns to inverting books, but now he is now no longer responsible for his actions. Kurt has been manipulated, and the relevant value has been nefariously implanted. However, according to the positive principles that we have examined, Kurt’s

springs and actions are authentic. This is because it would be true that Kurt had endorsed the value at an earlier time.

This type of case also affects negative principles. Suppose that Kurt was first brainwashed by the cult, then cured by librarians, and then came to adopt the value in question under his own steam. In this second scenario, Kurt would be responsible for inverting books. However, according to negative principles, Kurt's relevant springs and actions are not authentic. It would be true of Kurt that he had acquired the value under nefarious circumstances at an earlier time. The principles make no provision for the fact that at a later time Kurt acquired the value in a natural way.

The case of Fran having her values obliterated and replaced by equivalent values leads us to adopt a "type-centered" view of springs. The cases involving Kurt meeting positive or negative criteria due to events that are no longer relevant indicates that historical principles need to be modified in some appropriate way to accommodate such cases.

There is a straightforward solution to this problem. Principles must stipulate that only the most recent valid instance of reflection, acquisition, rejection, or implantation may be taken into account. In the first scenario, only Kurt's most recent instance of critical reflection is to be counted by the positive principle. Kurt had *rejected* and abandoned the value, so it is false that Kurt had (most-recently) supported, and he is not responsible, as expected. In the second scenario, only Kurt's most recent instance of acquisition is to be counted by the negative principle. It is false that Kurt's most recent acquisition was nefarious, and he is responsible, as expected.

### 3.5 A Collection of Principles

I propose that a historical theory should consist of a collection of principles, rather than a single principle. The principles in such a theory must address those scenarios in which a plausible change in the authenticity of a spring may occur. I provide a sketch of such a theory below. The principles listed are given in terms of values, but are intended to apply to any type of spring.

The first principle is a *manipulation* principle, and serves to define inappropriate modes of acquiring a value:

> (M) Any nefarious implantation of a value that is not already authentic for a person counts as the implantation of an inauthentic value.

The following two types of events or procedures count as nefarious manipulation. First, the direct alteration, construction, or destruction of certain neural structures that does not occur as a result of normal brain function. I leave open what counts as "normal brain function", but I exclude fantastic cases of neuro-surgical manipulation and cases involving ray guns. Such cases are cases of nefarious manipulation. Second, the inordinate exploitation of various human cognitive biases. I also leave open what counts as "inordinate exploitation", though I count cases involving heavy and prolonged doses of propaganda or "brainwashing" as being inordinate.

The next principle addresses the concern of manipulation that results in the endorsement or production of springs, and whether those springs are authentic. The principle asserts that manipulation is "*contagious*":

(C) Any process that was (or is) produced through nefarious manipulation or which produces values on the basis of inauthentic values (or springs) cannot produce authentic values.

The next principle is an "*absorption*" principle:

(A) Any inauthentic value that could reasonably have been abandoned by the agent before some time t but was not abandoned becomes authentic at that time t.

This principle is an analogue of the notion of "sheddability". It allows for the implicit endorsement (that is, absorption) of certain inauthentic springs. Like all other principles, it is sensitive to (C), in that the mechanism of shedding various springs (that is, of reducing their efficacy, perhaps until they are entirely ineffective) must not itself have been manipulated.

The following are principles of *endorsement* and *rejection*:

(E) A value that is endorsed by an agent becomes authentic.

I leave open exactly what counts as "endorsement". This is in effect the non-implicit form of (A). The principle is intended to accommodate cases in which a spring that arose through manipulation might be carefully considered and then endorsed by the person on the basis of authentic processes and springs. A person who has a strong disposition to like all books might be manipulated in such a way that she would have certain hospitable (and inauthentic) attitudes towards reading a certain book. But given the person's disposition, she might evaluate the book "under her own steam" and conclude that it is worth reading. In such a case, the relevant springs would count as authentic. Until endorsement occurs,

actions that arise on the basis of nefariously implanted springs are not authentic, with the exception of cases in which an agent implicitly endorses the relevant springs through (A) by failing to shed them.

(R) A value that is rejected by an agent becomes inauthentic.

Likewise, I leave open exactly what counts as "rejection". This principle is intended to apply to cases in which an agent has an authentic spring, but then comes to "reject" that spring on the basis of authentic springs. Consider the case of a smoker who freely chooses to become addicted to cigarettes, but later decides that it would be best to break the habit. This would, roughly, count as rejection of the desire to smoke. This rejection principle (R) would be sensitive to the absorption principle (A), in that a failure to take appropriate measures to abandon a certain value (or spring in general) entails that the value will again become authentic for that smoker. In other words, a failure to abandon a value when it is reasonably possible to do so is incompatible with the rejection of that value. Likewise, this principle (R) would also be sensitive to the agent endorsing (E) the value on the basis of various authentic springs (M, C). If the smoker comes to the conclusion that "having just one more cigarette would not be so bad", this would likely count as an endorsement of the relevant springs, and the smoker would be responsible if those springs led to smoking.

The following is a principle of *generation*, which specifies the initial status of springs with respect to authenticity:

(G) Any spring that arises is authentic at the moment that it arises.

Again, this principle is sensitive to (C). This principle (G) does not rely on the status of any past event: it simply affirms that, in the absence of manipulation, a spring is authentic. This places the theory in line with negative historical theories, since such theories have in common the fact that in the absence of manipulation, a spring is authentic.

The next principle is a more constrained formulation of a generation principle, which simply asserts that in a case like that of Suzie Instant, the agent is responsible.

> (S) Any spring that came into being at the same instant that a fully-developed agent came into being, or which "arose" through some suitably internal process not subject to external influence, is authentic.

The motivation for "suitably internal" is that if an instant agent is created at a moment, and in the next moments develops a complex variety of springs "under her own steam", then such springs would be authentic.

The principles here form a single historical theory of responsibility (call it MCAERG; I exclude S as being entailed by G). The theory, though relatively basic, takes into account the various issues noted and provides answers to the various cases presented in the second chapter. It is in line with the requirements for a historical theory of responsibility mentioned above.

Applying these principles to the Ann-Beth case yields the result that Beth is not responsible, because her relevant values were implanted in a nefarious way according to (M). Suzie Instant, however, is responsible according to (G).

## Conclusion

In the first chapter, I provided an account of action and of closely related concepts, including basic causation, intentionality, consequences, responsibility, springs of action, authenticity, and history. I gave, roughly, the following account. An action is a discrete event, with preceding causes and consequent effects. In acting, agents intend to bring about certain consequences. Various springs of action (such as desires, beliefs, values and dispositions) factor into a process of deliberation. The decision that arises through this process of deliberation is a direct action.

In the second chapter, I gave a critical account of various motivating cases and theories of moral responsibility, ranging from non-historical theories, to both positive and negative historical theories. I argued against the Drunk Driving case, and presented Mele's Ann-Beth case in favour of historicism, as well as McKenna's Suzie Instant case in favour of non-historicism. I set out Frankfurt's non-historical theory, and the historical theories of Mele, Fischer and Ravizza, and McKenna.

In the third chapter, I developed a basic framework of constraints on theories of responsibility in general. I set out a general argument against positive historical theories, and argued for the idea of springs as being more akin to types, rather than tokens. I then presented an initial sketch of a historical theory of responsibility, which I formulated as a set of principles.

## Bibliography


Cuypers, Stefaan E. 2004. “The Trouble With Harry: Compatibilist Free Will Internalism and Manipulation.” *Journal of Philosophical Research*, 29: 235–254.

Fischer, John Martin. 2000. “Chicken soup for the semi-compatibilist soul: Replies to Haji and Kane.” *The Journal of Ethics*, 4: 404–407.

Fischer, John Martin. 2004. “Responsibility and manipulation.” *Journal of Ethics*, 8: 145–177.

Fischer, John Martin. 2006. *My Way: Essays on Moral Responsibility.* New York: Oxford University Press.

Fischer, John Martin & Ravizza Mark. 1994. “Responsibility and history.” In *Midwest studies in philosophy*, 430–451.

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

Frankfurt, Harry G. 1971. “Freedom of the will and the concept of a person.” *Journal of Philosophy*, 68: 5–20.

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

Haji, Ishtiyaque. 1998. *Moral Appraisability.* New York: Oxford University Press.

Haji, Ishtiyaque. 2002. “Compatibilist views of freedom and responsibility.” In *The Oxford Handbook of Free Will*, ed. Robert Kane 202-28. New York: Oxford University Press.

Haji, Ishtiyaque. 2009. *Incompatibilism’s Allure.* Peterborough: Broadview Press.

Haji, Ishtiyaque & Cuypers, Stefaan E. 2004. “Responsibility and the Problem of Manipulation Reconsidered.” *International Journal of Philosophical Studies*, 12: 439–64.

Haji, Ishtiyaque & Cuypers, Stefaan E. 2007. “Magical agents, global induction, and the internalism/externalism debate.” *Australasian Journal of Philosophy*, 85:3, 343-371.

McKenna, Michael. 2004. “Responsibility and Globally Manipulated Agents.” *Philosophical Topics*, 32: 169–192.

McKenna, Michael. 2011. “Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories.” In the *Oxford Handbook of Free Will, Second Edition*, ed. Robert Kane 175-98. New York: Oxford University Press.

McKenna, Michael. 2012a. “Defending Nonhistorical Compatibilism: A Reply to Haji and Cuypers.” *Philosophical Issues*, 22: 264-280.

McKenna, Michael. 2012b. “Moral Responsibility, Manipulation Arguments, and History: Assessing the Resilience of Nonhistorical Compatibilism.” *The Journal of Ethics*, 16: 145–174.

McKenna, Michael. (forthcoming). “A Modest Historical Theory of Moral Responsibility.” *Journal of Ethics.*

Mele, Alfred. 1992. *Springs of Action*. New York: Oxford University Press.

Mele, Alfred. 1995. *Autonomous Agents*. New York: Oxford University Press.

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

Mele, Alfred. 2007. “Persisting Intentions.” *Noûs*, 41: 735–757.

Mele, Alfred. 2009a. “Moral Responsibility and Agents’ Histories.” *Philosophical Studies*, 142: 161–181.

Mele, Alfred. 2009b. *Effective Intentions*. New York: Oxford University Press.

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

Vargas, Manuel. 2006. "On the Importance of History for Responsible Agency." *Philosophical Studies*, 127: 351–382.