SELF-MOVEMENT AND NATURAL NORMATIVITY: KEEPING AGENTS IN THE CAUSAL THEORY OF ACTION

A Dissertation
Submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Philosophy

By

Matthew R. McAdam, B.A.

Washington, DC
October 16, 2007

Self-Movement and Natural Normativity:
Keeping Agents in the Causal Theory of Action

Matthew R. McAdam, B.A.

Thesis Advisors: Wayne A. Davis, Ph.D. & Margaret O. Little, Ph.D.

ABSTRACT

Most contemporary philosophers of action accept Aristotle's view that actions involve movements generated by an internal cause. This is reflected in the wide support enjoyed by the *Causal Theory of Action* (CTA), according to which actions are bodily movements caused by mental states. Some critics argue that CTA suffers from the *Problem of Disappearing Agents* (PDA), the complaint that CTA excludes agents because it reduces them to mere passive arenas in which certain events and processes take place.

Extant treatments of PDA, most notably those of Michael Bratman and David Velleman, interpret the problem as a challenge to CTA's ability to capture the role of *rational capacities* like deliberation and reflection in the etiology of human action. I argue that PDA admits of another interpretation, one that arises when we appreciate that the exercise of higher rational capacities in action presupposes possession of a prior lower-level capacity for *basic self-movement* – the power to initiate and control one's bodily behavior. Bolstering CTA so that it accommodates richer exercises of practical thought – as Bratman and Velleman do

– will not resolve PDA unless CTA already captures this basic agential power. Adequately responding to PDA, therefore, requires answering a question unaddressed by current responses: How do bodily movements caused by sub-agential items like mental states count as movements actively performed by the whole agent?

I argue that CTA can answer this question by adopting a normative account of the nature of self-moving agents. On this view, self-moving agents are *teleologically constituted*, meaning (1) their nature and proper function derives from their characteristic ends and aims, and (2) the nature and proper function of their parts depend on these ends and aims. Basic self-movement consists of movements caused by a sub-agential part whose own proper function is to generate behavior that constitutes or contributes to the pursuit of the agent's overall proper function. After showing how this picture applies to artifactual and collective agents (i.e., robots and teams), I extend the account to organic agents (human beings) by sketching a broadly Aristotelian picture of the nature of living things.

To Cheryl

TABLE OF CONTENTS

INTRODUCTION

0.1 The Causal Theory of Action (CTA)

The causal theory of action (CTA) – at least, the version of it that will concern me throughout this dissertation – is offered as an answer to the question, what happens when someone acts? This is the story it tells ('→' denotes causation):

Beliefs/Desires/Intentions → Bodily Movements

Thus, according to CTA, when someone acts, her motivating beliefs and desires and intentions[1] (what some philosophers call the agent's "motivating reasons") cause the agent's body to move in appropriate ways such that the movements constitute the action the agent intended.

CTA arises from the belief-desire model of action explanation. An agent's wanting something and believing something is translated into talk of causally efficacious mental

[1] Not all philosophers of action agree on the kinds of attitudes that need to go into CTA. For example, those who reject a Humean theory of motivation deny that desires must figure in the attitudes that move agents to act. Others deny that intentions are separate attitudes, claiming instead that an intention consists of a certain kind of belief/desire pair. I do not want to take a stand here on these issues, so my inclusion of all three mental states – beliefs, desires, and intentions – is meant to broadly illustrate the kind of mental states that could go into CTA; it should not be read as supporting any necessary conditions for CTA.

items, particulars endowed with the causal power to move limbs.[2] There is an important question about how exactly beliefs and desires cause limbs to move. Though no particular answer is entailed by the formulation of CTA, it is generally assumed that psychological attitudes cause limb movements by virtue of the fact that they are related to underlying physical states of the agent. The idea here is that all causal efficacy ultimately derives from the causal workings of physical causes and causal processes. The term 'physical' can be understood in more or less broad ways, and on its narrowest interpretation it refers to the micro-level of fundamental entities or properties, the ultimate constituents of the physical world. A broader, and more common, understanding of the physical is captured by reference to a "physicalist worldview" or the point of view of the "physical sciences" where these are generally taken to include physics, chemistry, biology and their various off-shots.

0.2 The Problem of Disappearing Agents (PDA)

CTA is sometimes accused of being guilty of the problem of disappearing agents (PDA):

> (PDA): In CTA the agent serves merely as an arena in which mental and physical events take place; the agent plays no active part in her movements.

[2] Hornsby 2004, 180.

As I argue in Chapter 2, the notion of an agent "playing an active part" in her movements is elusive. For now we can understand PDA as a problem of accounting for how the movements caused by parts of an agent like mental states or their neuro-physiological realizers can be equivalent to movements performed by the agent of whom they are states. Thus, from this point of view, to overcome the challenge PDA poses for the causal theory, we need an answer to this question: how does the causal efficacy of certain of an agent's parts transfer to the whole agent? Or, how can bodily movements caused by mental or neuro-physiological states be movements that are actively performed by the whole human being of whom they are states? The aim of this dissertation is to provide an answer to these questions.

The problem of disappearing agents results from CTA's reductionism, its exclusive focus on mental states and bodily movements in its analysis of human agency. Though these occurrences eventuate in movements of the agent's body, the fact that these movements owe their occurrence to certain inner events seems to challenge the notion that they are movements the agent herself is making. Why would a better description of the case not be that the agent is passively *being moved*, not that she is actively *moving*? The worry is that when an agent's mental states are understood as causally interacting inner items, they seem to become loci of causal powers that compete with the causal powers of the agent herself.

0.3 An Initial Objection

Before moving on, it is important at the outset to deal with a potential objection to this project. We can begin by noting that the thought behind PDA can be expressed with the something like the following question: Are not actions instances in which agents produce their bodily movements *themselves*, rather than instances in which their movements are produced by their mental states? Some feel the force of this question, others do not. The objection I am considering here is from those who do not– I'll call them Ardent Causal Theorists (ACT's). ACT's think that this question is nonsense. For, according to them, *what it is* for an agent to act intentionally *just is* for her movements to be caused in the right way by certain of her beliefs and desires. In other words, CTA articulates the *constitution* of an action, the *what it is to be an action*, i.e. bodily movements caused in the right way by appropriate mental states. Thus, ACT's interpret PDA's challenge to CTA as similar to, "How can it be that a heart attack is the sudden cessation of one's heartbeat?"[3] To ask this question is to betray a lack of understanding of the concept of a heart attack. There is nothing to say in response to this question other than, "Well, if you have to ask, then you just do not know what a heart attack is." In other words, there is simply no room for such a question. Similarly, one can ask what water is; but one cannot ask why water is H_2O. There is no reason *why* water has this chemical composition – it just does. In the same vein, actions *just are* bodily movements caused in the right way by beliefs and desires. ACT's think that to wonder about this is to simply not get CTA.

[3] I borrow this example from Wayne Davis.

ACT's unfairly characterize PDA. To see why let us stick with the heart attack example. There are many platitudes that go along with the concept of a heart attack. We can call the set of these platitudes the concept's *profile*. Part of the profile of the concept 'heart attack' is that it is something that causes severe chest pain, is correlated with numbness in the left arm (usually just in men), and tends to afflict persons who are overweight or have high blood pressure. So when folks were going about trying to figure out what heart attacks were, what they were doing was trying to determine what goes on in people with, say, high blood pressure that causes them to experience pains in their chest and numbness in their left arm, and that often times kills them. In other words, they were looking for what it was that answered to the profile of a heart attack. The event of one's heart suddenly ceasing to pump blood fits.

The following are some platitudes about human action: acting intentionally involves being in control of oneself and knowing what one's doing; acting intentionally is doing something for a reason; when we act intentionally we cause things to happen in the world; agents are responsible for their intentional actions because some of an action's results are traceable directly back to an agent. The profile of the concept of an intentional action is the general picture of intentional action that we get from these platitudes. Of course, this is not an exhaustive list of our intuitive thoughts about action, but it gives us enough to work

with.[4] So, the question to ask is, Does CTA fit the profile of the concept of an intentional action? Surely, *pace* ACT's, this question is not complete nonsense.

The following claim by Richard Taylor, one of the proponents of agent causation, is a clear expression of PDA's doubt about CTA's claim to capture the profile of intentional action:

> It is plain that, whatever I am, I am never identical with any such event … or state as is usually proposed [by CTA] as the "real" cause of my act.… Hence, if it is really and unmetaphorically true … that I sometimes cause something to happen, this would seem to entail that it is *false* that any event … or state not identical with myself should be the real cause of it.[5]

Now I am not suggesting we endorse Taylor's point (nor am I suggesting we should reject it). I do think, however, that it is a legitimate point that deserves a response. Why does not an agent have to be identical to whatever it is that causes the movements of her body that constitute her intentional actions? What is it about mental states (and/or their physical realizers) such that when they cause (in the right way) movements of agents we can say that the agent is moving herself and not just *being moved* by causes internal to her? How do mental-state-caused bodily movements get to count as intentional doings of the agent of whom they are states? How do we have to understand the nature of mental states in order for the movements they cause to redound to the agent? These are the questions I try to answer in this dissertation.

---

[4] There are also probably exceptions to each one of the platitudes I mentioned. Platitudes are weaker than necessary or sufficient conditions.
[5] Taylor 1966, 111.

0.4 A Sketch of the Argument

The first, negative part of my discussion begins with the complaint that philosophers who have made previous attempts to make CTA safe from PDA have misidentified the problem. I argue that this failure to properly address PDA stems the fact that contemporary philosophers of action tend to focus their attention on those aspects of agency that are unique to human or rational agents. We can explain this by invoking the idea of human beings as *rational animals*. Most current philosophical thinking about human agency focuses on the rational side of our nature and largely ignores the animal side. So attention is paid to rational mental processes like deliberation, reflection, and planning that precede uniquely human actions, while the bodily processes of initiating and controlling the movements that constitute such actions is generally ignored.[6] This is because the basic capacity for self-movement – the power to initiate and control one's bodily movements – is present throughout the animal kingdom. Black bears and giraffes have the capacity just as human beings do. Thus, it seems that if our concern is with what is special about human rational agency, then we need to turn our attention to something that is come and gone once the agent's actually moving herself.

This emphasis on our rationality over our animality affects the way philosophers interpret PDA's criticism that CTA leaves agents out of its story of action. For if the focus is on human beings, and if what is special about human agency is the capacity for practical rationality, then the claim the CTA fails to capture the role agents play in their actions

[6] An exception to this is Frankfurt 1988.

becomes the claim that CTA fails to capture deliberation, reflection, planning and the like. On this view, the agent "disappears" because her rational powers play no role in the story of action CTA tells. But notice that this takes for granted that CTA adequately captures the exercise of our capacity for self-movement. That is, the interpretation of PDA as a challenge to CTA's ability to give our higher rational faculties the right role in the etiology of human action rests on the assumption that there is no worry about an agent "disappearing" when beliefs and desires cause movements constituting actions that do not involve a prior engagement of these faculties – the kind of movements that we *and* non-rational animals can make. So, the idea that the movements constituting our actions are caused by our mental states is fine so long as the movements constitute the kind of actions that (so-called) mere brutes can do, activities like reaching for food or walking across the street.

But surely the complaint that agents in CTA fail to play an active role in their actions, i.e., that CTA describes passively being moved not actively moving, is just as plausible, if not more plausible, when applied to these simpler cases. The failure to capture exercises of practical rationality is a problem, perhaps even one adequately described in terms of an agent "disappearing." However, it is a more fundamental problem, one *literally* describable in terms of an agent "disappearing" or being left out, to be unable to give an account of what happens when someone acts that does justice to the fact that active bodily movements are due *to the agent herself* and not to her mental states. By the end of Chapter 2, I hope to have made a convincing case for an alternative interpretation of PDA, one that emphasizes CTA's threat to what I call basic self-movement.

The aim of the positive part of this dissertation, Chapters 3 and 4, is to demonstrate how CTA can overcome the charge that it cannot capture basic self-movement. This essentially involves developing a view of both self-movement and the nature of things endowed with the capacity for self-movement that allows movements caused by states (or, more generally, parts) of such things to count as movements performed by the thing as a whole. I begin developing this view by illustrating that CTA is an instance of a more general account of how things move themselves, what I call *the standard story of self-movement*. We can trace the standard story back to Aristotle's account of action in Book 3 of the *Nicomachean Ethics*, where he observes that action has its origin *within* the agent; action has an *internal* principle or source. Thus, the standard story's Basic Thesis (BT):

(BT) Self-movement is internally caused bodily movement.

BT informs more than our understanding of what happens when human agents act. I demonstrate this by initially discussing the standard story in the context of artifactual agents like robots and conventional/collective agents like teams. The standard story's generality as an account of self-movement rests on the fact that self-moving agents – artifactual, conventional, and organic – share a common structural feature, what I refer to as *constitutive teleology*.

To say that a self-moving agent is characterized by constitutive teleology, or to say that an agent is *teleologically constituted*, is to say that the agent as a whole has certain characteristic ends and aims, and that the agent's parts have certain proper functions that depend on these ends and aims. The possibility of self-movement, that is, cases in which

certain movements caused by a thing's parts count as movements performed by the whole, arises from the congruence between an agent's nature, as revealed by the agent's overall ends, and the nature of a certain part or mechanism within the agent.

As I argue in Chapter 3, we can get a grasp of the concept of constitutive teleology by looking at the collective actions of groups or institutions. Consider the following case: a quarterback throws a pass to one of his team's receivers; the receiver catches the ball and runs with it into the end zone; as a result of what these two individuals do, the entire team scores a touchdown and wins the game. Here we have a case of a complex whole (a football team) performing an action (scoring a touchdown) by virtue of parts of the team (individual players) doing certain things (throwing, catching, and running). What makes this kind of thing possible is the institution of a "normative space" determined the rules and standards – *the constitutive norms* – that determine the group's functional architecture – the various roles and offices that comprise the collective along with the varying forms of authority that go along with them.

I respond to PDA by extending this picture of collective agency to individual human agents, which requires demonstrating how the concept of constitutive teleology applies to human beings.[7] Here I rely on an Aristotelian account of living things. At the heart of the Aristotelian view is the idea that animals are constituted so as to attain the good endemic to its kind. This implies that the animal's parts, capacities, dispositions, tendencies to activity,

[7] Though I do not emphasize it here, I argue in Chapter 3 that constitutive teleology also characterizes the constitution of artifactual agents like robots. Thus, the normative view of self-movement I develop has the virtue of applying to three paradigmatic examples of self-moving agents: robots, teams, and animals.

etc., are structured and organized so as to serve as instruments for the attainment of the good of whole animal they comprise. We can restate this in modern parlance, in terms that are perfectly naturalistic and consistent with evolutionary biology. For another way of speaking of an animal's form is in terms of an animal's species kind; and another way of speaking of an animal's good is in terms of the life-cycle – the process of development, self-maintenance, and reproduction – of this kind of creature. Thus, combining Aristotelian and evolutionary language, animals have a teleological constitution determined by their life-form (species) that enables the creature to fully participate in the life-cycle of its kind, which is paradigmatically equivalent to developing into a healthy, mature, full-functioning specimen.

Now here is the connection between the point about collective actions and the Aristotelian account of living things: an animal's (Aristotelian) form is akin to a political constitution or a set of rules for a game – it bestows an identity on the parts that make it up, and assigns to certain parts particular roles or functions that enable the whole to do certain things. Both group agents like teams, and, on the Aristotelian account, individual agents like human beings, have teleological constitutions. That is, they both embody a normative structure that determines the identity and function of their parts. The argument is that just as individual players on a team can do things that amount to the whole team of which they are a part doing something, so too there is an element of a human agent's teleological structure that can function in a way that amounts to the whole of which it is a part moving itself.

In Chapter 4, I argue that the element of the human agent endowed with this power is the faculty of desire. Desires are understood as essential to self-movement because they are understood as essentially connected to the good of the whole. The function of the desiderative capacity is to track the relevant good within the context of choice and action, and to put the agent in pursuit of it. Movements caused by the faculty of desire constitute movement of the whole agent for two reasons: first, the agent is constituted so as to attain its good; and, second, the faculty of desire is a part of an agent whose proper function is to track the good and move the agent toward it. Putting these two points together, we can say that movements caused by desires are caused by an agent's *nature*, that which makes the agent that embodied creature she is. So, movements caused by a part of the agent whose purpose is to track the good count as self-movement because they are movements caused by the agent's essential nature. In the end, I argue that embracing the Aristotelian account of living things makes available to us a way of understanding the nature of action-involving mental states like desires that reveals how CTA does not result in the agent's disappearance.

Finally, I want to make clear the ambitions of this dissertation. In the discussion that follows, I am not aiming to provide a complete theory of agency or provide a complete defense of the causal theory of action. My more limited aim is to develop a compelling response to what I take to be the major challenge facing the causal theory. This means that I am not calling into question the truth of CTA. Rather, I am articulating a view about the nature of self-movement and self-moving agents that we must see as in the background of CTA's story about what happens when someone acts.

ONE

The Causal Theory of Action

1.1 Introduction

In this chapter I introduce the causal theory of action (CTA), discuss Davidson's influential argument in its favor, and sketch two different interpretations of the causal theory. After articulating these dual conceptions of CTA, I demonstrate how they both fall prey to the problem of disappearing agents (PDA). The overall project of this dissertation is to develop a response to PDA on CTA's behalf. I begin with a mundane example of intentional human action.

While sitting at her desk Mary reaches over to the window and slides it open. She does this because she wants to let some fresh air into her office. If someone asked Mary why she opened the window she could explain it by making reference to this desire for fresh air. Given this information we could also sensibly ascribe to Mary the belief that opening the window would be a way to get some fresh air. We can also sensibly assume that had she wanted, say, to get a bowl of cereal instead of some fresh air, then she would have behaved

differently. Thus, her beliefs and desires *exert an influence* on her; they *make a difference to* her behavior. These italicized locutions are examples of common ways we have of talking about the relationship between mind and action, or between what people think and what they do. What this brings out is that there is a *causal component* to our understanding of commonsense psychological attitudes like belief and desire.

In fact, we can test this. Suppose that just before reaching over to open the window Mary expresses to you her desire for some fresh air. Now imagine that you convince Mary that there is a toxic cloud of poisonous gas floating outside her window. If Mary's rational, and let us assume she is, then she clearly will not open the window now. Why? Well, your talk of the toxic cloud influenced her thinking, it led her to change her mind. Prior to talking to you she thought opening the window was perfectly safe, now she thinks it is dangerous. This change in belief led to a change in desire. Instead of wanting to open the window, Mary now wants to cover it with the plastic wrap her clever government said would come in handy in a case like this.

1.2 The Causal Theory of Action (CTA)

These commonplace thoughts about the effect of our beliefs and desires on our behavior are captured in the causal theory of action (CTA). Here is David Velleman's articulation of CTA:

> There is something that the agent wants, and there is an action that he believes conducive to its attainment. His desire for the end, and his belief in the action as a means, justify taking the action, and they jointly cause an intention to take it, which

> in turn causes the corresponding movements of the agent's body. Provided that these causal processes take their normal course, the agent's movements consummate an action, and his motivating desire and belief constitute his reasons for action.[8]

Michael Smith provides a similar account:

> Actions are bodily movements that are caused and rationalized by an agent's desire for an end and a belief that moving her body in the relevant way will bring that end about.[9]

Finally, here is what Jaegwon Kim has to say about the causal theory:

> the possibility of human agency evidently requires that our mental states – our beliefs, desires, and intentions – have causal effects in the physical world: in voluntary actions our beliefs and desires, our intentions and decisions, must somehow cause our limbs to move in appropriate ways.[10]

According to Kim, mental states and events bring about the movements that constitute our action. Our capacity to perform basic actions, and therefore the foundation of our agential capacities, is the power of our desires (or our beliefs and desires) to cause our limbs to move. Kim makes this explicit when he claims that having our mental states

> cause our limbs to move in appropriate ways … is how we manage to navigate around the objects in our surroundings, find food and shelter, build bridges and cities, and destroy the rain forest.[11]

Reverting to the example of Mary opening the window, we can trace the reasoning behind CTA. We start with the claim that Mary's action of opening the window is explained by her wanting some fresh air and by her thinking she can get some by opening the window. The standard story converts the claim that Mary's attitudes are explanatorily relevant because they

---

[8] Velleman 2000, 123.
[9] Quoted in Hornsby 2005, 2.
[10] Kim 1998, 31.
[11] Kim 2005, 9.

affected her behavior into the claim that Mary's attitudes caused the bodily movements that constitute her opening the window. Mary's wanting something and her believing something is translated into talk of causally efficacious mental items, particulars endowed with the causal power to move limbs. So what happens when Mary opens the window because she wants air and thinks that the way to get some is to open the window is her desire for air and her belief that opening the window fulfills that desire jointly cause the movements of her body (her reaching) that constitute her act of reaching out and opening the window.[12]

There is an important question about how exactly beliefs and desires cause limbs to move. Though no particular answer is entailed by the formulation I have given of the standard story (or the causal theory more generally), it is generally assumed that the causal story will be filled out in physicalist or materialist terms, meaning that psychological attitudes cause limb movements by virtue of the fact that they are related to underlying physical states of the agent. The idea here is that all causal efficacy derives, ultimately, from the causal workings of physical causes and causal processes. The term 'physical' can be understood more or less broadly, and on its narrowest interpretation it refers to the micro-level of fundamental entities or properties, the ultimate constituents of the physical world. A broader, and more common, understanding of the physical is captured by reference to a "physicalist worldview" or the point of view of the "physical sciences" where these are

[12] I am assuming here and throughout the rest of the dissertation that, unless I claim otherwise, the causal chains from attitudes to bodily movements in the examples of action I consider "take their normal course", thereby setting aside worries about deviant causal chains.

generally taken to include physics, chemistry, biology and their various off-shoots. Tyler Burge explains the basic physicalist (or materialist) intuition as follows:

> There is certainly reason to believe that underlying our mental states and processes are physical, chemical, biological, and neural processes that proceed according to their own laws. Some such physical processes are probably necessary if intentional (or phenomenal) mental events are to be causes of behavior.[13]

What it means exactly to say that the physical "underlies" the mental is not entirely clear, and there are various ways of understanding the relation. The minimal physicalist commitment is that the mental supervenes on the physical, meaning that there can be no change at the mental level without a lower level physical change. In other words, "mentality is at bottom physically based … there is no free-floating mentality unanchored in the physical nature of objects and events in which it is manifested."[14] Most contemporary philosophers of mind and action go further and support some form of token-identity theory, according to which each mental particular (event or property instantiation) is identical to a physical particular. In the case of action, this kind of view allows for an alignment of the story of mental causation told by CTA with the underlying neuro-physiological story that can be told any time there is activity involving human bodies. Acceptance of token-identities between mental and physical particulars accounts for the tendency of many philosophers, when the topic is agency, to speak in terms of bodily movements being caused by causally interacting "inner"

[13] Burge 1993, 103.
[14] Kim 1998, 14-15.

states of agents, where this is meant to refer to the intra-cranial entities that are the province of the cognitive sciences.[15]

1.3 The Naturalist Background to CTA

CTA appeals to many contemporary philosophers because it seems to *naturalize* agency by resolving a tension between a broadly scientific understanding of the world and our commonsense thinking. Velleman explains this as a tension between the idea of "agential origin" at the heart of our concept of agency, and a "naturalistic conception of explanation" that is endemic to a scientific worldview. "Our concept of … human action," he writes, "requires some event or state of affairs that owes its occurrence to an agent and hence has an explanation that traces back to him."[16] On the other hand, "our scientific view of the world regards all events and states of affairs as caused, and hence explained, by other events and states, or by nothing at all."[17] "Actions," John Bishop concludes, "thus seem to involve something that natural science does not recognize."[18] This tension between the requirements of agency and a naturalistic picture of the world's causal workings sets the agenda for the project of developing a causal theory of action:

> Much of recent philosophy of action has been motivated by the reductionist project of finding a place for human agency in a naturalistic picture of the world. In such a

[15] This literal construal of talk of "inner" states does not necessarily apply to all physicalists, but it is clear that this is what many philosophers of mind mean when they talk in this way. This is perhaps most evident in Fodor's work, and other philosophers similarly committed to the representational theory of mind and the concomitant language of thought thesis. See Wilson 1995, 152-53. For criticism of talk of "inner" states see Collins 1986 and Steward 1997.
[16] Velleman 2000, 127. See Bishop 1989, 2.
[17] Velleman 2000, 129. See Yaffe 2000, 121.
[18] Bishop 1989, 31.

> picture, actions are not caused by a primitive entity called "the agent" or "the conscious self", but by states and events that are amenable to a scientifically respectable description.[19]

CTA attempts to provide "a psychological reduction of what happens in rational action", an account of agency that only "alludes to states and events occurring in the agent's mind."[20] The causal theory reduces an agent's moving her body to an agent's mental states causing her body to move. For example, if Hoover intentionally waves to me from across the quad, CTA claims that Hoover's arm is moving back and forth because some relevant mental state of Hoover, such as his desire to say hello, is causing his arm to move that way. Note that this is a metaphysical reduction, not a conceptual or analytic one. CTA does not claim that the concept "moving a limb" reduces to the concept "mental states causing the limb to move." Clearly these express difference thoughts. What CTA does claim, however, is that whenever it is true that an agent is intentionally "moving a limb" this is true by virtue of the fact that the agent's "mental states are causing the limb to move". Another way of expressing the reductive point is this: *what it is* for an agent to intentionally move a limb *just is* the agent's mental states causing the limb to move (in the right way); or, an agent's intentionally moving *is constituted by* the agent's mental states causing the movements. We can ascribe the source of an agent's movements to the agent herself, in other words, when the movements are caused in the right way by the agent's motivating beliefs and desires.

I characterized the naturalistic conception of explanation in terms of a commitment to the view that states and events are the only things that can stand in causal relations, and

[19] Schroeter 2004, 650.
[20] Velleman 2000, 130.

that therefore they are the only things we can appeal to when explaining the occurrence of natural phenomena. One might reject this characterization of naturalism by appeal to the fact that we often point to things that are neither states nor events when giving explanations of natural occurrences. Vivid examples of this are when we say that an atomic bomb caused the destruction of Hiroshima; or that the planes that hit the Twin Towers caused them to collapse.[21] The point is that there seems to be nothing non-naturalistic or anti-scientific about such explanations, suggesting that any form of naturalism worth taking seriously would not preclude them. This criticism is surely correct about the way we talk about events in the world. Most naturalistically-minded philosophers, however, would reply that the surface grammar of sentences that predicate causal powers of objects like bombs and planes is misleading, for it masks an underlying event-causal story that grounds the explanatory power of these statements. The naturalist does not deny that we give these sorts of explanations, or suggest that there is anything spooky about them. What the naturalist does deny is the claim that the bomb, *as opposed to the event of the bomb's exploding*, caused the destruction of Hiroshima. The naturalist maintains that there *is* something strange about this, and insists that acceptance of such a statement is not at all part of scientific naturalism. So the naturalist claim is not that we do not predicate causal powers of agents when explaining what people do: the claim is that we should not read the metaphysics of causal explanation off the surface grammar of such claims.

[21] I owe this objection to Wayne Davis.

The naturalist's avoidance of appeals to things or substances as irreducible causes is a version of a general methodological principle that characterizes scientific investigation, namely, the method of decompositional or reductive analysis. This is the practice of reductively analyzing an entity in terms of its constitutive components, and then giving an account of its behavior that appeals to the workings of its parts. There tends to be some resistance to analyzing human agency in the same way. In his influential article "Free Will as Involving Determination and Inconceivable without It", Hobart chides some philosophers for suffering "a want in the analytic imagination" when they turn their attention to human action. "We have been accustomed," he writes, "to think of [a person's] activities as the way in which, as a whole, [she] naturally and obviously behaves." When we exercise the "analytic imagination", however, we "realiz[e] that the component parts of a thing or process, taken together, each in its place, with their relations, are identical with the thing or process itself."[22] An example of the analytic imagination at work in contemporary philosophy of mind is Lycan and Dennett's "homuncular functionalism". As Lycan explains,

> we view a *person* as a sort of corporate entity which corporately performs many immensely complex functions – functions of the sort usually called "mental" or "psychological". A psychologist who adopts [this] AI-inspired methodology will describe this person by means of a flow chart, which depicts the person's sub-personal agencies and their many and various routes of "access" to each other … which enable them to cooperate in carrying out the purposes of the containing "institution" or organism that that person is."[23]

Though CTA itself does not take the same reductive functionalist shape of the view Lycan articulates, it is still an example of the decompositional strategy. For it breaks down an

[22] Hobart 1934, 2.
[23] Lycan 1981, 28.

agent's capacity for intentional movement in its environment into a story about the causal interactions among mental states and limbs. And this is, in fact, an important step in the process of the kind of "flow chart" strategy characteristic of much current philosophical and scientific thinking about the mind.[24]

1.4 Davidson's Defense of CTA

A central concern in contemporary philosophy of action has been articulating the difference between actions – things that agents do – and mere movements – things that simply happen to agents. This is the light in which most philosophers of action have read Wittgenstein's question, "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?" As Velleman explains, "the difference between my arm's rising and my raising it is supposed to illustrate the difference between a mere occurrence involving my body and an action of mine" (1). This twofold distinction between actions and mere bodily movements has structured the context in which CTA has developed.

The distinguishing mark that separates mere bodily movements from actions is that the latter, unlike the former, can be understood as active expressions of an agent's rational capacities. That is, a human action is a nexus of activity and reason. As Elizabeth Anscombe stressed in *Intention*, human actions are happenings "to which a certain sense of

[24] Note that CTA – the view that actions are bodily movements caused by mental states – is not necessarily wedded to this kind of reductionism. The reason for this is that it is not essential that CTA take a physicalist form. A Cartesian dualist can be a causal theorist, and surely dualism is as recalcitrant to naturalistic reduction as any. As should be clear from the discussion above, however, for the purposes of my discussion I am assuming that causal theorists are physicalists who assume that actions and the mental states that cause them are "realized by" or are "identical to" or "supervene on" (etc….) the physical.

the question 'Why?' is given application; the sense is of course that in which the answer … gives a reason for acting." [25] Certain rationalizing links must be in place between what an agent does and the reasons that motivated her to do it in order to properly see her behavior as action. I will call this the *rationalizing links requirement*.

The question of whether reasons are causes, or whether rationalizing explanations are causal explanations, exercised a great many philosophers a couple of decades ago. Without question, the view that reasons are causes, that rationalizing explanations are a species of causal explanation, came out on top of this debate and it has for the most part gone unquestioned ever since. Donald Davidson's article, "Actions, Reasons, and Causes" is generally credited as establishing the prominent status the causal theory enjoys in contemporary philosophy.[26] The rationalizing links requirement is at the heart of Davidson's argument. He argues that the rationalizing link that is in place when beliefs and desires cause actions is secured via the causal process through which actions come about, and he challenges his non-causal opponents to articulate an alternative conception of the relation between an action and the reason for which it was performed – one that does not rely on any spooky non-natural relation, but which secures the appropriate rationalizing link.

Let me take a moment to give some background to Davidson's argument. Contemporary philosophy of action begins with Anscombe's *Intention*, and Anscombe and the many philosophers who followed in her wake were heavily influenced by what they saw

[25] Anscombe 2001, 9.
[26] Davidson 1980.

as the conceptual analysis of agential concepts in Wittgenstein's later work.[27] These writers emphasized the ways in which rational explanations of human actions differ from the kinds of causal explanations we give for other events in the natural world. Rational explanations, they argued, are normative, and they work by fitting the action being explained into a certain pattern of rule following or a broader social practice. The anti-causalists also specifically argued that the relation between psychological attitudes like intentions and desires, and the actions motivated by these attitudes is logical, not causal; rationalizing explanations of action are conducted in a different "logical space" from causal explanations. To explain an agent's action in terms of her reasons and psychological attitudes is simply to do something very different from picking out antecedent states and events whose occurrence resulted in the action.[28]

---

[27] The Wittgensteinian view found its greatest expression in the "sea of little red books" to which Davidson refers in a later article. This is a reference to Routledge & Kegan Paul's series *Studies in Philosophical Psychology*, edited by R.F. Holland, a series that many credit as establishing the central role of philosophy of mind in contemporary philosophy. The series includes early works by Norman Malcolm (on dreaming), Anthony Kenny (on moral psychology), and Alistair ManIntyre (on the unconscious).

[28] To do full justice to the anti-causalist arguments, it is important to see them in the context of mid-20th-century debates about whether the natural sciences competed with or subsumed the so-called human or social sciences, e.g. economics, psychology, sociology, and anthropology. If one could argue that such disciplines had different ways of explaining phenomena than those practiced by the natural sciences, and argue that they are none the worse for this, then one has mounted a good case for their legitimacy as sources of knowledge about human nature. Unfortunately, this context is now largely overlooked, leaving the philosophers of "the little red books" with hardly any sympathetic readers in contemporary philosophy. For outside this context their work can appear to be mere quibbling about what we do when we explain actions. The real nature of their concern, I think, can only be unraveled by acknowledging the many ways in which we use the concept of 'cause' and the implications that follow regarding the nature of persons when we settle on a particular sense as capturing the explanatory connection between reasons and the actions they explain. I think a charitable interpretation of the Wittgensteinian anti-causalists sees them as aimed at combating scientistic tendencies in philosophy the influence of which were just starting to be felt at the time they were writing. For recent defenses of non-causal accounts of reasons and rational explanations of action see Schueler 2003, Sehon 2005, and Wilson 1989.

At the heart of the anti-causal position is the "logical connection argument". It begins by noting one of Hume's points about causation, namely that part of the concept of causation is the notion that cause and effect exist separately, that is, that the causal connection is a contingent one. As Melden, a prominent non-causalist, explains, "the very notion of a causal sequence logically implies that cause and effect are intelligible without any logically internal relation of the one to the other."[29] On the other hand, the relation between intentions or desires and the actions they motivate is a non-contingent logical relation; an agent's action and the desire or intention that motivated it cannot be rendered intelligible independently of each another, for the intentional content of these mental states is about the resulting action. Therefore, so the argument goes, when we advert to an agent's intentions or desires in order to understand her action we are not pointing to antecedent conditions that are contingently connected to the action, as they would be if they were the action's causes. Rather such explanations render an agent's behavior more conspicuous by providing a meaningful interpretation or re-description of it. For example, Vivian's *wild gesticulations* make sense when I learn that she is *dancing*, just as Todd's initially odd *sounding-off* is heard in a whole new light when I realize he is making the Muslim *call to prayer*. So coming to know of someone's intentions or desires is not to discover what caused the agent's action, rather having an account of the content of these mental states allows us to see the purposes in a person's movements, to understand the pattern in their behavior.

[29] Melden 1961, 82.

In "Actions, Reasons, and Causes" Davidson devastated the "logical connection argument" by drawing a distinction between events and their descriptions. Logical relations, he argued, are intensional; they exist between descriptions of events. What this implies is that the existence of a logical relation between a statement of the content of one's intention, belief, desire, etc., and a description of one's action does not preclude those psychological attitudes from being the action's causes. Whether event X is a cause of event Y has nothing to do with how we account for or describe the events – the latter's a linguistic, *de dicto* affair, while the former is an extensional, *de re* matter. Logical and causal relations, therefore, are perfectly compatible and can exist simultaneously between the same events. So reasons certainly can be causes, or at least the "logical connection argument" does nothing to preclude this. This is because there may also be true descriptions under which the reasons, understood as mental states and events, do in fact stand in causal relations to the bodily movements that constitute the resulting actions.

Davidson also attacked the non-causalist's appeal to the context in which an action takes place as a way of explaining an action. We see this in Melden's example of the driver signaling a turn with his arm.[30] Melden argues that we can explain why the driver raised his arm by paying attention to the situation in which he performed this action, namely, at an intersection in the midst of making a turn. He argues that in this context, and in light of a background familiarity with the social norms involved in driving a car, one can explain the driver raising his arm by coming to see it as a signaling of a turn. We can say that the

[30] Melden 1961, 99.

driver's intention in raising her arm is to signal a turn, but when we advert to a mental item in this way in order to explain an action we are not pointing to something antecedent to the signaling that brought about the arm raising. Rather, to cite the intention is simply another way of saying what it is the driver's doing *by raising her arm* – she is signaling a turn.

> In response, Davidson begins by accepting the point about redescription:
>
> when we explain an action, by giving the reason, we do redescribe the action; redescribing the action gives the action a place in a pattern, and in this way the actions is explained.[31]

He argues, however, that agreeing to this settles nothing about the nature of the explanation we give when we give an agent's reason for acting. Just because giving the reason has the effect of placing the action in a wider context of, say, a certain social practice does not preclude the reason giving explanation from being causal. "Talk of patterns and contexts," he writes, "does not answer the question of how reasons explain actions, since the relevant pattern or context contains both reason and action."[32] The suggestion here is that understanding the explanation as causal *does* provide an answer to the question of *how* the agent's reason explains her action. Davidson's complaint is that the non-causal view offers no alternative to this account of how it is that citing the right mental items – which Davidson agrees can have the effect of re-describing and contextually situating the action – actually succeed in explaining what the agent does. What the appeal to patterns and practices does is establish that in certain situations an agent has a reason to do something, e.g. raise his arm, but what the appeal does not do is establish that this reason that the agent

[31]Davidson 1980, 10.
[32] Davidson 1980, 10.

has really is *the reason* for which the agent acted. "Central to the relation between a reason and an action it explains is the idea that the agent performed the action *because* he had the reason."[33] And this is what Davidson thinks cannot be captured without an appeal to causality. The appeal to causation can provide the needed "analysis of the 'because' in 'He did it because …' where we go on to name a reason."[34]

This last point has come to be known as Davidson's "master argument." One cannot overstate the influence of Davidson's master argument in turning the tide of the philosophy of action from the anti-causal view to the causal theory. Mele refers to the master argument as "Davidson's Challenge" to philosophers who oppose the causalist view that reasons explanations are causal:

> If you hold that when we act intentionally we act for reasons, provide an account of the reasons for which we act that does not treat (our having) those reasons as figuring in the causation of the relevant behavior (or, one might add, as realized in physical causes of the behavior)! [35]

Though there have been some attempts, this challenge is generally regarded as still outstanding.

1.5 An Ambiguity in "Causal Theory of Action"

At the beginning of "Actions, Reasons, and Causes" Davidson announces that his aim is to defend the view that explanations of actions that cite an agent's reasons (what Davidson

[33] Davidson 1980, 9.
[34] Davidson 1963, 11.
[35] Mele 2000, 279-80

refers to as rationalizations) are a species of causal explanation. We can express this thesis about the nature of action explanations as

(CE) Rationalization is a species of causal explanation.

Soon after he identifies CE as his thesis Davidson introduces the notion of an agent's primary reason. He explains that an agent's primary reason comprises a pair of attitudes, namely, "some sort of pro attitude toward actions of a certain kind" and a belief that "his action is of that kind."[36] Relying on the notion of a primary reason Davidson restates his thesis in these terms: "the primary reason for an action is its cause". We can express this thesis about the nature of reasons for action as

(C) Reasons are causes.

CE and C are two common slogans used for expressing the causal theory of action. What is the relation between CE and C? Initially it seems that Davidson thinks C is simply a restatement of CE, another way of expressing the same thought or making the same claim. This is evident in the fact that he refers to the claim expressed in C as part of a reformulation of the view the rationalizations are causal explanations. Later on, however, Davidson distinguishes between the extensional relation of causation that obtains between particulars and the intensional relation that obtains between sentences, or the truths or facts that they express, in a true causal explanation. This distinction between causation (or causal relations) and causal explanation suggests a difference between CE and C. CE explicitly

[36] Davidson 1980, 3-4.

makes a claim about causal explanations, while C, on the other hand, is a claim about causes or causation.[37]

Davidson's dual understanding of the causal theory of action is reflected in the way current philosophers express allegiance to the causal theory. It is clear from what different philosophers say that those who call themselves causal theorists do not all accept the same view. A conspicuous way of understanding their differences is in terms of the distinction between CE and C. If we look at the literature we can detect a clear difference between those philosophers who understand the causal theory of action as a thesis about reasons explanations – Davidson's *rationalizations* – and those who interpret the theory as a story about the events and processes that produce or bring about actions.

Allow me to cite at some length a sampling of some statements regarding the causal theory of action to illustrate this point. I will start with philosophers who understand the causal theory along the lines of C. For example, Jaegwon Kim expresses his understanding of the causal theory of action in the following terms:

> We are agents, agents who can deliberate, form intentions and action plans, and act so as to realize what we intend. This means that in performing an action, our decisions and intentions, and our desires and beliefs, cause our limbs to move in such a way that things around us get appropriately altered and rearranged.[38]

John Bishop defends the causal theory in similar terms:

> According to the *Causal Theory of Action*, actions consist in behavior that is caused by appropriate mental states – mental states that make it reasonable for the agent to perform behavior of that kind. What we think of as *agents* doing thing, it is

[37] See Strawson 1985.
[38] Kim 2000, 5.

> suggested, is actually a matter of *certain of their mental states* causing those things to occur.[39]

The latest example of a defense of this way of understanding the causal theory of action is Berent Enç's recent book *How We Act*. He articulates the causal theory of action as follows:

> [A] causal theory of action…has as its central thesis the proposition that an act consists of a behavioral output that is caused by the reasons an agent has…that consist of the beliefs and desires of the agent…. On this thesis, actions are defined as changes in the world that are caused by mental states.[40]

These statements by Kim, Bishop, and Enç express an understanding of the causal theory of action best captured by C. Their statements are not about what we are doing when we explain actions, but rather about how the limb movements that constitute our actions come about. While a claim about the explanations of action is about the epistemology of actions, about our knowledge and understanding of them, these statements are best interpreted as metaphysical claims about the nature of action. In other words, they are statements about the kinds of causal chains involved in the production or generation of actions, or at least the movements that constitute them. This is to interpret talk of an agent's beliefs and desires in explanations of their actions as referring to items in the realm of particulars that play a role in bringing about the event that was the agent's action. Now, Bishop and Enç go a step further than Kim does by suggesting that mention of these causal processes is part of an analysis or definition of the concept of action. But all three agree that the causal theory of action is a metaphysical view about the nature of actions, in the sense that it is a theory about the causal process involved in the generation of actions. Reasons, an agent's

[39] Bishop 1989, 2.
[40] Enç 2003, 2.

motivating beliefs and desires, are particulars that play a central role in this causal process which is why this metaphysical version of the causal theory of action can be expressed as C, i.e. reasons are causes. This is not to say that proponents of this causal theory action do not also think that we can cite these reasons in order to explain someone's action, and that when we do so we are giving a causal explanation. While a commitment to a causal theory in the form of C does not entail acceptance of CE, proponents of C do in fact tend to endorse CE as well.

Some philosophers, however, only endorse CE. This is a version of the causal theory of action that does not purport to make claims that tell a causal story about what happens when someone acts. Rather, these philosophers understand the causal theory to be a view about the nature of the reason-giving explanations we rely on to understand what people do. For example, Jennifer Hornsby argues that, "the causal reality of belief and desire is just their causal-explanatory reality". She asks:

> Why should acknowledgement that we say something about what she believed and desired in causally explaining why she did what she did lead us to accept the existence of anything that 'the cause of her action' stands for?[41]

Hornsby is expressing skepticism about the notion that when we cite an agent's beliefs and desires in explaining her behavior what we are doing is mentioning causally efficacious particulars that caused – in the sense of produced or brought about – her action. Rather, she argues, we should "deny that the causal-explanation view relies on the idea of discrete things

[41] Hornsby 1997, 134.

combining (interacting?) in the production of action."[42] So Hornsby is not saying that human actions are somehow outside of the causal domain. She is claiming that our commonsense psychological explanations of action, while they are causal explanations, do not work by identifying the items that played a role in the production of the movements that constitute an action. She argues that this misplaces that subject matter of rationalizing explanations by mistakenly focusing on the event that is the action instead of the agent who performed the action. It is to hear questions about why an agent Ø-ed as asking for an account of why a Ø-ing occurred.

> When we seek an 'action explanation', one question we usually want answered is 'Why did she do such and such a thing?' We may agree that actions are events without supposing that this question is equivalent to 'Why was there an event of such-and-such a kind?' Asking why *a* Ø-ed, we hope to learn something about *a*, the person; but if we asked why *a*'s Ø-ing occurred, *a* might not be the subject of concern at all.[43]

William Child also endorses a version of the causal theory that restricts itself to the nature of reasons explanations:

> [A]s I am using the phrase, a causal theory of action is simply a theory which says that explaining actions by giving the agents' reasons for doing them is a mode of causal explanation.[44]

Child is explicit about the fact that other philosophers think there is more to the causal theory, and it is clear from his remarks that what he has in mind is the kind of view Kim and Bishop discuss. He writes:

---

[42] Hornsby 1997, 135.
[43] Hornsby 1997, 134.
[44] Child 1994, 99.

> Some causal theories of action have included other claims. In particular, some writers associate the causal theory of action with the view that S's Ø-ing was an action if and only if it was caused … by an intention, or by a belief and desire.… But someone who claims that reason explanation is a form of causal explanation can remain neutral on that point….[45]

Unlike Hornsby, though, Child does think that rationalizations are a species of causal explanation because they explain why something happened or why some event occurred. In fact, he thinks that this is what makes it the case that rationalizations are causal; for what all causal explanations have in common, he argues, is "that they are all explanations of the occurrence or persistence of particular events or circumstances, or of general types of event or circumstance."[46] He does, however, agree with Hornsby's rejection of the need to appeal to "causally interacting internal entities" in order secure the causal nature of rationalizations. "We can have a causal theory of action," he writes, "without holding that propositional attitudes are causes of behavior."[47] That is, we can say that explanations of the form "she Ø-d because she believed that *p*" are causal explanations without saying "if S Ø-d because she believed that *p*, her belief that *p* must have been an item causally implicated in producing her behavior."[48]

---

[45] Child 1994, 99. Thus, unlike Hornsby, Child believes that acceptance of a version of the causal theory of action that is best captured by CE – "rationalization is a species of causal explanation" – does not entail a stand on whether the causal theory also involves C – "reasons are causes".

[46] Child 1994, 100. I see no reason why we cannot accept Hornsby's point while also going along with Child's view that rationalizations are causal explanations because they explain why something happened. Hornsby's right to say that we want to learn something about the agent when we request an explanation for her behavior. But, at the same time, discovering the relevant information about her just is a way of explaining why the event that was her action occurred. In other words, by finding out why she did what she did, we learn why her doing it occurred.

[47] Child 1994, 125.

[48] Child 1995, 125.

At this point I have put two different versions of the causal theory of action on the table, which correspond to Davidson's two different statements of the causal theory (C & CE). There is a metaphysical version of the causal theory that gives an account of the causal role an agent's psychological attitudes play in the generation of the movements that constitute her action. There is also an explanatory version of the causal theory of action that is silent on the metaphysical question of how actions come about, or of what happens when someone acts. This a causal theory of action explanation, one which says that rationalizations are causal explanations. These two versions of the causal theory see beliefs, desires, and other practical attitudes as playing very different roles in relation to action. On the metaphysical version, these attitudes have causal *efficacy*, meaning they are caught up in the processes that result in the action. On the explanatory view, these attitudes have causal *explanatory relevance*, meaning we can advert to them when giving a causal explanation of an agent's action.

The focus of this dissertation is on what I referred in the Introduction as "the problem of disappearing agents," a challenge that several prominent philosophers of action have raised against CTA. In the next section I will introduce this problem and discuss the possibility of embracing the causal explanatory version of CTA as a way of avoiding the problem.

1.6 The Problem of Disappearing Agents Introduced

As I explained above, CTA is the result of an attempt to respond to what Velleman refers to as "the fundamental problem in the philosophy of action," namely, "finding a place for agents in the explanatory order of the world."[49] Agency's recalcitrance to positioning within the natural explanatory order is due to the concept of "agential origin" inherent in our commonsense concept of action. "Our concept of full-blooded human action," Velleman explains, "requires some event or state of affairs that owes its occurrence to an agent and hence has an explanation that traces back to him." The problem, however, is that a "naturalistic conception of explanation", one dictated by "our scientific view of the world," rejects the idea of tracing an explanation of the occurrence of any states or events in the world back to anything other than states and events. The result is that "any explanation of human action [must] speak in terms of … occurrences, because occurrences are the basic elements of explanation in general."[50] When we remind ourselves of the obvious point that an agent is not an occurrence, a state or event, the conflict between agency and the natural world becomes evident.

CTA attempts to resolve the conflict between agency and the prevailing natural scientific method of explanation by articulating a "psychological reduction of what happens in rational action."[51] That is, CTA offers a theory of agency that only "alludes to states and

[49] Velleman 2000, 127,
[50] Velleman 2000, 130.
[51] Velleman 2000, 124.

events occurring in the agent's mind."[52] In this more austere picture of agency, agential origin is accounted for in terms of the occurrences that supposedly make it up. Its austerity stems from the way it proceeds without expanding the naturalist event-causal ontology, that is, without relying on a fundamental appeal to a different metaphysical kind of thing, an agent (more generally, a substance), other than the kinds recognized by naturalism. If this reductive, event-causal story can be told about human action, then apparently philosophy of action's fundamental problem has been solved.

Note, however, that if CTA, in its attempt to psychologically reduce human agency, cannot produce a surrogate of agential origin out of states and events, then it fails to reconcile agency with the strictures of scientific naturalism. For if the reduction of agential origin (a phenomenon whose explanation requires appeal to something beyond mere occurrences) to the causal interactions among states and events actually amounts to its *elimination*, then the standard story is no longer a story about human agency. (At least, that is, so long as we accept that agential origin is an essential part of human agency.)

Several recent prominent critics of CTA, e.g. Harry Frankfurt, David Velleman, Michael Bratman, have leveled this eliminitivist charge against the standard causal theory of action (CTA). Though there are differences in the particular formulations of their complaints, a point I will get to in Chapter 2, they all share the thought that there is something inadequate about CTA, that there is something important about human agency

[52] Velleman 2000, 130.

that is gone missing. And what they all agree goes missing is the agent herself. As Enç recently put it:

> Conceiving of action as the result of a chain of events that are causally connected removes the *agent* from the picture altogether. Being the conduit for a casual chain is a *passive* affair; being an agent is being *active*. The former cannot possibly exhaust what is involved in the latter.[53]

The problem with CTA is that "in this story reasons cause an intention, and an intention causes bodily movements but nobody – that is, no person – *does* anything."[54] In cases of genuine agency, on the other hand, "the *agent* is the source of, determines, directs, governs the action and is not merely the locus of a series of happenings, of causal pushes and pulls."[55] When an agent acts, Velleman writes, "an intention is formed by the agent himself, not by his reasons for acting [and] the agent then moves his limbs in execution of his intention: his intention does not move his limbs by itself."[56] CTA, therefore, fails to capture the essential feature of agency, namely, the fact that when he acts an agent is *active* as opposed to passive: he is not *being moved*, he is *moving*. Instead of being seen as active beings that bring something about through their agency, Frankfurt suggests that CTA's psychological reduction transforms agents into "locales in which certain events happen to occur."[57] "Psychological and physiological events take place inside a person," Velleman

---

[53] Enç 2003, 3-4.
[54] Velleman 2000, 123.
[55] Bratman 2001, 311.
[56] Velleman 2000, 122.
[57] Frankfurt 1988, ix.

writes, "but the person serves merely as the arena for these events: he takes no active part."[58] This is *the problem of disappearing agents* (PDA).[59]

Now it may appear that PDA only arises for the metaphysical interpretation of CTA, what I referred to above as C. The reason for this is that C is the version of CTA committed to beliefs and desires being causally efficacious inner states of agents that bring about the movements that comprise their actions. The explanatory version of CTA (CE), on the other hand, is not committed to this view. According to CE, beliefs and desires are attitudes that figure in causal explanations of actions, but not because these attitudes played a role in bringing about an agent's movements. Beliefs and desires are causal-explanatory because they figure in certain counterfactual supporting explanations the account for why something happened, namely, an agent's action.

While proponents of CE like Hornsby and Child may succeed in putting forward an interpretation of the CTA that does not demand embracing the picture of mental states that seems to result in PDA, this does not necessarily suffice for overcoming PDA. The reason for this is that we can restate PDA in a way that does not make any reference to causally efficacious mental states. For recall that PDA is the worry that the agent herself plays no

---

[58] Velleman 2000, 123.

[59] Horgan (2007) focuses on a problem that seems similar to PDA, what he calls the "agent exclusion problem." Horgan's problem involves an apparent incompatibility between CTA and agency that arises from reflection of the phenomenology of acting: "You experience your arm, hand, and fingers as being moved *by you yourself* – rather than experiencing their motion as … being caused by you own mental states" (187). PDA, on the other hand, involves an apparent incompatibility between CTA and agency that arises from reflection on the concepts of "agential origin" and "mental-state-caused bodily movements." Thus PDA remains even if Horgan's problem is solved. For we may be satisfied in rejecting phenomenological challenges to CTA, while still hankering for an answer to the question of how bodily movements caused by mental states count as movements generated by the whole agent of whom they are states.

role in her movements when these movements are the result of the causal powers of sub-agential items. Notice, though, that we can simply skip over beliefs and desires and go straight to what, according to a broadly physicalist view of the mind, would be their neuro-physiological realizers. In other words, it is not essential to PDA that the items endowed with causal powers that threaten the agent's own causal powers are mental. Rather, they can be neuro-physiological. The problem has the same structure either way – states of or parts of the agent, not the agent herself, cause the movements of an agent's body.[60] And notice that, while there may be room to deny the inner-causes model of mental states as CE does, there seems to be no room – short of embracing dualism – of denying the causal role our neuro-physiology plays in generating the movement of human bodies.

In the discussion that follows, I will generally interpret CTA along the lines of C – the version of CTA committed to beliefs and desires as inner items endowed with causal powers. At times, however, I will make explicit that we can also raise PDA's challenge in terms of the sub-personal neuro-physiological realizers of our mental attitudes.

1.7 Conclusion

In this chapter I introduced the causal theory of action and traced two different interpretations of the theory. I also introduced the problem of disappearing agents. In the next chapter, I focus on PDA in more depth. After clarifying the various things proponents

[60] Of course, there is a reading of this claim that should not be denied by anyone. This is to interpret it as the claim that in order for the agent herself to make any movements, to physically do anything, her parts must cause things to happen. The reductionist claim is that *what it is* for the agent to move herself is nothing other than these causal happenings among her parts.

of PDA could mean when they refer to an agent "playing an active role" in her behavior, I examine three of the most influential treatments of PDA in the work of Harry Frankfurt, Michael Bratman, and David Velleman. I argue that a version of PDA survives their responses, one that challenges CTA to answer the question of how movements caused by sub-agential items like mental states can count as movements made by the whole agent. In chapters 3 and 4 I sketch an answer to PDA's challenge that relies on embracing an Aristotelian conception of the nature of living things and a traditional conception of desire as a faculty that aims at the good. My aim is not to suggest that the view I offer is the only way for CTA to respond to PDA; rather, my goal is to make the Aristotelian view I suggest sound sufficiently plausible and compelling to warrant serious consideration as a way of thinking about the causal theory of action.

Two

The Problem of Disappearing Agents

2.1 Taking an Active Part?

PDA claims that the problem with CTA is that agents do not play a role in the story of action the theory tells. Recall Velleman's complaint that in CTA's picture of agency, "physiological and psychological events take place inside a person, but the person serves merely as an arena for these events: he takes no active part."[61] A problem with understanding PDA and generating a response to it is that this notion of a person "taking an active part" in her action or being sufficiently "involved" with her behavior is elusive. What exactly does it mean to say that an agent plays an active role in her actions?

We can get a handle on PDA's notion of "taking an active part" in our behavior by considering various categories of bodily movements, and asking what it might mean to say that the agent is or is not actively participating in them.[62] In this section I articulate three different categories of bodily movements (a "spectrum of agential involvement") that I will use to help us understand PDA.

---

[61] Velleman 2000, 123.

[62] Let me clarify some terminology. In what follows, talk of "playing an active role in", "being involved with", "participating in", and other similar locutions refer to the same relation between an agent and her movements, i.e. the one PDA claims CTA fails to secure.

Here are the categories of agential involvement that will be my focus:

1. *Robust Human Action*: rationally motivated behavior that results from conscious deliberation and choice, e.g. eating a plate of veal after reflecting on the relevant moral issues.
2. *Mundane Activity*: intentional movement not preceded by a conscious exercise of practical reason, e.g. reaching out and opening the window because you feel warm.
3. *Involuntary movement*: bodily movements neither initiated nor controlled by the agent, e.g. having a seizure.

Though we do not (yet) have a clear understanding of PDA's talk of "playing an active role in" one's behavior, it seems intuitive that these three cases represent varying degrees of this phenomenon. We can agree, I think, that as we go from 1 to 3 the agent becomes less involved in her behavior; she plays less of role in what she is doing.

This spectrum of agential involvement is clearly not exhaustive, nor is it intended to be. We can easily imagine various other cases that implicate the idea of an agent's active involvement or role in her behavior that do not neatly fit into the three above categories. However, as will become clearer below, what concerns me are two specific ways in which we can understand agential involvement. The first involves the extent to which an agent's *rational capacities* for practical thought are involved in her behavior. The second has to do with the involvement of an agent's *capacity for basic self-movement*, i.e., the power to initiate and control movement one's bodily behavior. I will argue that discussions of PDA have focused

exclusively on the first issue, leaving the question of whether CTA tells the right story about the much more basic phenomenon of self-movement untouched.

So as to forestall possible objections to my choice to focus on the three particular forms of agency I articulated above, I will start by saying a few things about various other examples of behaviors that I do not consider.

First, there are ways of extending the upper and lower limits of the spectrum. Starting with the upper end, there are cases of human agency that are even more robust than the first case. For example, we may want to say that cases involving more long-term planning than that found in 1 are cases in which the agent is even more involved with her behavior. I can think of two reasons for this. First, planning is a sophisticated form of deliberation that takes time and concentration, and therefore it does not seem to be the kind of activities we can imagine going on "off-line", as it were, or without the agent's involvement. Second, actions that require planning, those that are more temporally extended, tend to play a greater role in our lives and our identities (a point Bratman emphasizes). At the other end of the spectrum, we may want to say an agent is even less involved in her behavior than she is in the seizure case when, say, she is blown about by the wind. The difference here is that one could at least point to an internal cause of the movements comprising the seizure, and thereby at least minimally implicate the agent's body. That is, we can at least say in the seizure case that the movements were caused by internal mechanisms inside the agent. This, of course, does not allow us to say that the movements comprising her seizure are ones she is intentionally or voluntarily *performing*, because the

mechanisms involved in their production are not those that subserve the capacity to initiate and control self-movement.

So much for the upper and lower limits to the spectrum. We can also imagine intermediate cases of agential involvement that fall between the steps I articulate. For example, I do not mention a form of agency that likely falls somewhere between steps 2 and 3. I'll refer to this kind of behavior as *mere voluntary movement*. Here I am thinking of mindless activities like idly tapping one's foot while reading or drumming one's fingers on a table while listening to a lecture. While this kind of behavior is similar to what we find in 2, to the extent that it is voluntary movement that is not preceded by an exercise of practical reason, it does not quite rise to the level of mundane activity (as I am understanding it). One notable difference between these two forms of intentional movement involves their level of sophistication, and, concomitantly, the degree to which we can say the movements are performed without thinking about them. Mundane activity can be something like getting up off the coach and opening a window. This qualifies as more than mere voluntary movement because it involves things like competently navigating one's environment and having some awareness of, say, being warm or needing fresh air. Mere voluntary movement, on the other hand, need not involve skill or any conscious awareness.

There are a couple of other interesting cases to consider.[63] The first is the phenomenon of weakness of will. Where do weak willed actions fall on my spectrum? The answer is that they do not fall in any one place. The reason for this is that some weak willed

[63] Maggie Little and Wayne Davis (respectively) suggested these cases.

actions might involve a great deal of planning and reflection, and thus qualify as more than basic self-movement. At the same time, the diminished power of self-control exhibited in instances of weak willed action surely does affect in some sense the notion of an akratic's involvement in her behavior.[64] So the mundanely active agent may be less involved in her behavior than the akratic when the focus is on the engagement of some of her rational powers, such as the power to plan and deliberate, while she may be more involved in terms of other powers like self-control. That is, the agent in the midst of mundane activity may be doing something that, *were* she to reflect about it, she would end up endorsing as the thing to do. The akratic yet more robustly active agent *is* reflecting on what she is doing – hence her being more involved in her behavior in the sense of the engagement of rational capacities – but she is conflicted about her behavior and cannot fully control her execution of it.

The second kind of case I do not consider involves the consequences of one's actions and the extent to which one intentionally brings them about. For example, I reach for one of two switches on the wall in an effort to turn on the light. I hit the wrong one and accidentally turn on the fan instead. Reaching for the switch is a case of mundane activity in which it seems appropriate to say I am playing an active role; that is, I am not sleepwalking, nor do I unintentionally bump into the wall. This suggests that if all had gone as planned and I had turned on the light, then I would have played an active role in bringing about the

[64] Cases of weakness of will also raise questions beyond those relating to PDA's concept of "playing an active role", questions involving responsibility and the appropriateness of praise and blame. I take these to be separate, further questions that go beyond those about agential involvement. Consideration of these topics is beyond the scope of this discussion. Suffice it to say that robust involvement, on the scale I articulate, does not entail that one qualifies for full moral responsibility or normative evaluation.

light's being on. When things do not go as planned, and I turn on the fan instead of the light, perhaps we can say that I did not play a fully active role in the fan's being on.

The above discussion illustrates that PDA's concept of "playing an active role" in one's behavior is not univocal. As we have seen, an agent can be more or less actively involved in her behavior to the extent that she plans ahead, that her actions have their intended consequences, that she is not acting akratically, or that she is not mindlessly fidgeting. And, no doubt, there are other ways in which the notion of agential involvement can arise. None of these will be my concern, however. I am skipping over these cases to avoid muddying the already murky waters that talk of "playing an active role" in one's behavior gets us into. What is at issue between PDA and CTA can be addressed more clearly by emphasizing the two particular ways of understanding agential involvement I mentioned above – the full engagement of rational faculties and the basic capacity for self-movement. In order to narrow in on these two understandings of agential involvement, let me now turn to the three cases of agency that comprise what I have been calling my spectrum of agential involvement.

The first case, *robust human action*, is an example of behavior that issues from the full exercise of the agent's capacity for practical rationality. The agent conducts and guides the deliberative process prior to the behavior, and she initiates and controls the resulting movements that constitute her action. It is clear that the agent is fully involved in robust human action; there is no sense in which she is left out or fails to play a role in what she is doing. When it comes to PDA's concept of playing an active role in one's behavior, this is

an easy case. Notice, however, that most of the actions we perform involve little, if any, careful thought or prior deliberation. For this reason we cannot draw the lesson from the consideration of robust human action that what it is for an agent to play an active role in her behavior just is for her movements to come about via conscious thought about how to act. Otherwise, most of our voluntary behavior will not qualify as movements we actively perform.

During any given stretch of time, if we are acting, it is likely we are participating in *mundane activity* – activities like turning on a faucet, walking a flight of stairs, steering one's car out of the driveway, or pushing a cart around the grocery store while scratching items off one's shopping list. If the agent plays less of a role or is less involved in this kind of behavior (compared to robust human action), then this must be because she does not fully exercise her powers of practical reason.[65] But just like robust human action, mundane activity is initiated and controlled by the agent; and this allows us to say that here too the agent is playing an active role in what she is doing. For the absence of conscious deliberation or reflection from an agent's action does not render her a helpless bystander to the movements her body makes. Even movements we make without consciously thinking about them can be movements we *perform*, things we *do*.

[65] Note that the absence of conscious practical reasoning from the immediate causal process leading to basic self-movement does not imply the complete absence of such thought from the broader causal history of our actions. There are many instances of common behavior that were not always part of one's repertoire of basic self-movement, behaviors that only become seamlessly incorporated into one's life after careful consideration of the reason (moral or otherwise) that support so acting.

Thus, if we want to say that the agent is not as fully involved in mundane activity (or non-robust action), we need to be clear about what it is we are claiming. We do not want to say that the lack of involvement registered at this level (when compared to robust human action) renders an agent's behavior defective or less than ideal. While stepping back and carefully considering how to act is appropriate in some situations, it is neither appropriate nor desirable much of the time. Someone who, in the regular course of things, devoted much thought to whether, for example, she should first shop for the items at the top of her list or those at the bottom, or whether she should use her left or her right hand to turn on the faucet, is not thereby functioning as an ideal agent. We may want to say that agents who *never* exercise their deliberative capacities are deficient as agents (though we should not assume this is even possible). But we do not want to say that there is anything deficient about having these capacities and not always (fully) exercising them. As Williams taught us, there are times when acting automatically in response to one's situation is exactly what is called for. In these situations, thinking about how to act amounts to having "one thought too many."

It is important to keep in mind that, in distinguishing between robust agency and mundane activity in terms of the involvement of rational faculties, I am not suggesting that, given the relative infrequency of instances of robust agency in our daily lives, we get along most of the time without the use of our rational capacities. As rational animals our lives are imbued with rationality; as Robert Brandom observes, we "live and move and have our

being in the space of reasons."[66] As a result, our reaching out and grabbing a banana, peeling it, and eating it is different from a monkey's doing so; and these differences can be cashed out in terms of the norms governing human eating and the role eating plays in a human life. Part of maturing as a human being consists in getting better at grasping these norms and structuring one's behavior in accordance with them. So while mundane activity is characterized by the lack of full engagement with one's rational faculties prior to and during action, this does not imply that this kind of workaday behavior is thereby completely divorced from these faculties. The fact that all of our behavior takes place within a structure of norms guarantees that all of our actions involve rationality. What is important for my purposes, though, is that the issue of whether one's behavior conforms to social norms does not track the issue of whether one's playing an active role in one's behavior. The explicit engagement of our rational powers does, I argue, track this difference, which is why I have focused on this.

The difference between robust human action and mundane activity has to do with the extent to which an agent's practical reason plays a role in the immediate causal history of her behavior. The line separating these two cases, however, is not the line that determines whether an agent plays an active role in her bodily movements. Intuitively, it seems agents play an active role in both robust human action and mundane activity because the movements that constitute both forms of behavior result from an exercise of the agent's capacity to initiate and control her bodily movements. I will refer to this package of

[66] Brandom 1994, 5.

capacities as an agent's *capacity for basic self-movement.*[67] This is not to deny that the distinction between robust human action and mundane activity marks a genuine difference in forms of agency. The full engagement of practical rationality makes an important difference to the nature of the movements that result from it; and we need not have a problem describing this difference in terms of an agent being more involved in or playing a more active role in them. However, if an agent's body is moving, but these movements are not a result of an exercise of her capacity for self-movement, then what is going on with her does not qualify as agency at all. Though the agent's moving, she is not performing her movements or bringing them about herself.

I am arguing that the distinction between robust human agency and mundane activity does not track the important difference between cases when agents are voluntarily moving themselves and cases in which agents are simply being moved (by either external or internal causes). In order to draw *this* line, the line between cases in which an agent plays an active role in her movements and those in which she does not *tout court*, we need to compare robust human action and mundane activity with the third form of movement in the spectrum of agential involvement, *involuntary movement*. The seizure case is a clear example of bodily movements in which the agent plays no role at all. This is because seizures *afflict* those who suffer from them. They are events that *happen to* someone, things one *undergoes*. Missing

[67] In packaging the capacity to initiate movement with the capacity to control movement, I am obviously skipping over the question of whether these capacities necessarily go together. Though there may in fact be cases of creatures with only one or another of these capacities – no doubt it is possible whether or not it is actual – for the sake of my discussion I am only going to focus on those who have them both. I think the best interpretation of PDA's notion of "playing an active role" in one's behavior involves, at minimum, the two capacities.

from the causal etiology of the person's bodily movements is any engagement at all with either her reason or with her basic capacity for self-movement. What this means is that, unlike the other two cases, the third case it is not an instance of *agency* at all. Things are *happening* here, but nothing's *being done*. This case is aptly described in terms of the agent "disappearing", and not of just playing a less active role in her behavior.

2.2 Two Important Distinctions

We can draw a couple of conclusions from the above considerations. By comparing robust human action and mundane activity we see that we can be actively involved in our behavior without deliberating or fully engaging our rational capacities before or while acting.[68] What secures this involvement is what is common to both cases, namely, the exercise of an agent's basic capacity for self-movement, that is, the fact that the agent initiates and controls the movements that constitute her action. In other words, in both robust human action and mundane activity the agent is actively *moving*, not passively *being moved*. This is why, unlike the seizure case, the first two cases are examples of *agency*. They are both ways of being *active* instead of *passive*. The first two cases, though they differ in the degree to which an agent's higher-order rational faculties are involved, are both examples of *activity*, movements that the agent *performs*. In the case of involuntary movement, the person is completely passive; none of the relevant movements are traceable to her as initiator or source. Here we can say the

[68] If the proponent of PDA disagrees, then her standard for what counts as agency is too high. It would imply that most of what we do fails to qualify as agency. It would also imply that only rational creatures are agents, thereby making automatons out of most of the animal kingdom. Both of these implications are implausible.

agent does not play any role in her bodily movements; her body's moving but she is not moving it. Thus, in order to understand PDA and the concept of 'playing a role in one's behavior' endemic to it, we must be careful not to confuse what separates robust human action and mundane activity *from each other* with what separates *both* forms of agency from involuntary movement.

The examination of the three kinds of bodily movements has yielded two distinctions:

1. Activity vs. Passivity (cases 1 & 2 vs. case 3)
2. Robust Agency vs. Non-Robust Agency (case 1 vs. case 2)

The first distinction drives a wedge between agency and mere bodily movements, while the second distinction drives a wedge between kinds of agency. To be an agent *simpliciter* is to have one's movements fall on the left-hand side of the first distinction, to have the basic capacity for self-movement – the power to initiate and control the movements of one's body. This is what allows us to say that there can be non-rational and inanimate agents. Obviously, cats and the Mars Rover are not capable of robust agency, but again this more sophisticated form of action does not set the minimal criterion for having the capacity for purposive activity.

Many philosophers would reject this notion of agency as the exercise of the basic capacity for self-movement as too thin. One reason for this is that a common assumption among many philosophers of action and moral psychologists is that agency is irreducibly normative, meaning it essentially involves being sensitive to and responding to reasons. That is, one common assumption is that the two capacities mentioned above, reason and the

power to initiate and control action, must be one and the same. Even richer versions of this rationalistic understanding of agency claim that to be an agent is to have a status that qualifies a person as an apt candidate for praise and censure (Strawson's "reactive attitudes"). On this account of agency, not only must a creature be sensitive to reasons in order to be an agent, the creature must also be situated within a community of similarly endowed rational agents.[69] We need not deny that this normative conception may capture one way we have of thinking about agency, an understanding that brings thoughts of autonomy and responsibility in its wake. However, we should reject it as an account of the minimal criteria of what it is to be an agent.

The most compelling reason for rejecting this rationalistic conception is that it over-intellectualizes agency to such an extent that it narrowly restricts the class of agents to developed, well-functioning human beings. But human beings are not the only creatures who *do* things. As Harry Frankfurt explains:

> agency is not unique to human beings or even to humans together with those various less evolved animals that may be regarded as also capable of some mode of practical reasoning. The difference between a creature that is actively in control of its own movements and a creature that is being moved passively by forces over which it has no control is familiarly instantiated even among species too primitive to engage in rational or deliberative thought. When they are active rather than passive, the members of those species function as agents even though processes of practical reasoning cannot plausibly be attributed to them.[70]

---

[69] Note the requirement of communal situatedness is not to be understood as something in addition to or over and above the reasons sensitivity requirement. The reason for this is that, on this view, being reasons sensitive, i.e. being a rational being, itself necessarily requires communal situatedness. See Brandom 1994.
[70] Frankfurt 2002, 90.

Frankfurt illustrates this with the example of a spider. If we administer an electric shock to the spider its legs will move, yet the spider will not be moving itself; the spider is simply moving, just as the limbs of a tree move when blown by the wind. In this case, Frankfurt says, the spider is "passive with respect to the movements of its legs."[71] When the spider is busily spinning a web or devouring its prey, on the other hand, it is active with respect to the movements – initiating and controlling them. Though this behavior lacks any connection to rational capacities, it still amounts to activity, movements that the spider performs rather than undergoes. The same reasoning applies to inanimate agents as well. There is a difference, say, between an engineer raising a robot's arm and the robot raising it.[72] The fact that the thin conception of agency, which characterizes agency in terms of the basic capacity for self-movement, captures these other cases of active individuals speaks in favor of rejecting the normative, rationalist view as setting the minimal criterion of agency.

Let me briefly recap at this point. The examination of PDA's notion of taking an active part or playing an active role in behavior resulted in two distinctions – the active/passive distinction and the robust/non-robust distinction – that we can use to frame our evaluation of PDA's criticism of CTA. Corresponding to the two distinctions are two senses in which an agent can play an active role in her behavior. The active/passive distinction is the basis upon which we determine whether someone's bodily movements count as agency. The criteria here are whether the agent exercised her basic capacity for self-movement in producing the movements that constitute her action. If the agent's

[71] Frankfurt 1988, 58.
[72] See Hauser 1994.

movements arise from an exercise of this capacity, then we can say that the agent is playing an active role in the behavior. On the other hand, she is not involved in the process if her body's moving but she is not moving it. Once this determination has been made and we have movement that qualifies as agency, then the second distinction applies. In applying the robust/non-robust distinction, we ask to what extent did the movements issue from the agent's rational capacity for deliberation and reflection. If the movements are the immediate upshot of practical reasoning, then we have robust agency. What it means to say that the agent is playing an active role in her behavior here is that it is an expression of an engagement of those powers and capacities that make her rational. Thus, to say an agent's not playing an active role in non-robust agency means that the behavior does not immediately implicate the agent's rationality.

The question now is, when PDA criticizes CTA for failing to allow an agent to play an active role in her behavior, is the complaint that CTA does not capture exercises of the basic capacity for self-movement? Or, is the complaint that, while CTA does capture basic self-movement, and thereby tells a sufficient story about mundane activity, it fails as an account of robust human action? As I argue in the following section, the focus in the contemporary debate over PDA has been on CTA's ability to account for robust human action. The charge is that we need more than beliefs and desires in our motivational psychology in order to get the exercise of rational capacities into CTA's picture of action. As I go on to argue, focusing on the more sophisticated forms of agency that humans are

capable of leaves untouched the more fundamental question of whether CTA captures the basic capacity for active self-movement.

2.3 Current Approaches to PDA

We can now return to the problem of disappearing agents (PDA) and the causal theory of action (CTA). At the outset of this discussion, I turned to David Velleman for an articulation of PDA. His complaint is that if agency works the way CTA says it does, then an agent is merely an arena in which certain movement-generating events take place. As a result, the agent "takes no active part" in her behavior. In a similar vein, Michael Bratman notes that in cases of genuine agency "the *agent* is the source of, determines, directs, governs the action," whereas CTA positions the agent as "merely the locus of a series of happenings, of causal pushes and pulls."[73] Inspired by Harry Frankfurt's hierarchical conception of the will, Bratman and Velleman have both developed sophisticated versions of CTA with an eye to repairing the shortcoming they see aptly expressed by PDA. I use the label *representative reductionism* to refer to the sort of theory they develop because the additional, higher-order mental states they add to CTA are meant to stand-in for the agent herself in the causal process leading to and resulting in intentional movements.

In the discussion that follows, I trace the development of PDA in the contemporary literature from Frankfurt's hierarchical account of the will to Velleman and Bratman's attempts to use this account to fortify CTA against PDA. After laying out the essentials of

[73] Bratman 2001, 311

Velleman and Bratman's views, I return to the two distinctions that fell out of the discussion of PDA's notion of playing an active role in one's behavior. I argue that while their views may enable CTA to better capture the robust/non-robust distinction, they fail to acknowledge the possibility that the PDA arises at the level of the more fundamental distinction between activity and passivity. If CTA cannot secure an agent's involvement in her behavior at this basic level, then no matter what kind of sophisticated causal theory of robust agency one develops (as Bratman and Velleman do), it will never get the agent actively involved in the story of action CTA tells.

2.3.1 Frankfurt

A leitmotif of Harry Frankfurt's writings on action is the important role the capacity for critical self-evaluation plays in the lives of human agents. Frankfurt thinks this reflexive evaluative capacity is an essential characteristic of personhood.[74] He describes this capacity in terms of the *hierarchical structure of the human will*. The structural components of the will's hierarchy are an agent's desires, and a desire's position on the hierarchy – its "order" – is determined by the object of the desire and by the kind of reflection that goes into its formulation and maintenance. The object of a *first-order desire* is the world, in the sense that it is a desire to pursue a certain course of action, to make the world a certain way. Having first-order desires seems to require no reflective capacities at all. Moving up the hierarchy, the object of a *second-order desire* is an agent's first-order desires; they are desires about desires.

[74] Cf. Korsgaard 1997 and 2002

"Someone has a desire of the second order," Frankfurt explains, "when he wants a certain desire to be his will."[75] The capacity to formulate second-order desires is constitutive of the capacity to critically monitor and exercise control over the desires that move us to action. As a result of having the power to form second-order desires, "when a *person* acts, the desire by which he is moved is either the will he wants or a will he wants to be without."[76] Having this ability opens up to a person the question of what her will is to be; that is, a person is someone for whom the nature of her will is an issue, a problem to be solved or handled, struggled with or ignored. An agent has the will she wants when she is moved to action by first-order desires that are the object of some of her second-order desires.

Frankfurt illustrates his hierarchical picture of the will by considering cases of defective agency, instances in which an agent feels torn between conflicting motivational impulses and is unable to exert rational control over her behavior. He imagines an unwilling addict who "hates his addiction and always struggles desperately, although to no avail, against its thrust":

> He tries everything he thinks might enable him to overcome his desires for the drug. But these desires are too powerful for him to withstand, and invariably, in the end, they conquer him. He is an unwilling addict, helplessly violated by his own desires.[77]

This unfortunate agent experiences two troubling psychic conflicts, one between two first-order desires and the other between a first and second-order desire. The first conflict results from the fact that the unwilling addict both wants to and does not want to take the drug

---

[75] Frankfurt 1988, 16.
[76] Frankfurt 1988, 19.
[77] Frankfurt 1988, 17.

simultaneously. His first-order motivational attitudes are pulling him in two directions. The second conflict arises because the unwilling addict is not ambivalent about which of these desires he wants to be the one upon which he acts. It is the desire to avoid the drug, not the desire to take it, that the he "wants to be effective and to provide the purpose that he will seek to realize in what he actually does."[78] By having this second-order desire to avoid the drug the unwilling addict *identifies* himself with this desire, thereby disassociating himself from the urge to light up. If the force of this second-order desire does not effectively render the desire to avoid the drug motivationally efficacious, then, according to Frankfurt, the addict is *alienated* from the desire that moves him and from the behavior it generates. This is not to say that the unwilling addict does not get what he wants; he does, after all, desire the drug. But this is decidedly not the kind of desire he wants to be acting on. A drug addict is not the kind of person he wants to be. Acting on the desire for the drug conflicts with a conception of himself he is come to value upon reflection, one that he is struggling to attain.

Reflection on the case of the unwilling addict case, and the behavior brought about by motives with which the agent does not identify herself – the phenomenon of *alienated agency* – leads to PDA's talk of an agent not playing an active role in her behavior. Frankfurt himself describes the unwilling addict's situation in similar terms: "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 … against his will that this force moves him

[78] Frankfurt 1988, 18.

to take it."[79] Bratman and Velleman pick up on this language and use it as a catalyst for the development of a richer causal theory of action, one equipped with the resources to respond to cases of alienated agency.

2.3.2 Bratman and Velleman

One lesson to draw from Frankfurt's view is that any adequate philosophical theory of human agency must contain the resources to account for our capacity for critical self-evaluation. Frankfurt accounts for this capacity by appeal to higher-order desires. However, as Gary Watson argues, Frankfurt's hierarchical theory does not preclude the possibility that agents will not also experience the same sense of alienation they do from some of their first-order desires from their second-order desires as well.[80] And if this is a possibility, then identification with action-motivating desires via endorsement by higher-order states cannot guarantee non-alienated action. It is this limitation of Frankfurt's hierarchical view to which Bratman and Velleman are directly responding. Unlike Frankfurt, however, they are both

[79] Frankfurt 1988, 18. It is important to note that Frankfurt is not actually raising the problem of disappearing agents (PDA) here. This is because (unlike Bratman and Velleman) he is not concerned with CTA and the project of developing a reductive theory of human agency. His aim, rather, is diagnostic. The hierarchical picture of the will is offered up as an illustrative way of articulating what is distinctive about the forms of agency available to rational creatures like us, not as a complete metaphysics of agency. Therefore, the fact that, like CTA, Frankfurt's hierarchical view of the will only appeals to mental states and events is not the basis for foisting on Frankfurt the ambition of giving a naturalized theory of action. In other words, we misread Frankfurt's work if we conclude from the fact that Bratman and Velleman both rely on Frankfurt in the service of reductive naturalism that this is what Frankfurt himself was trying to do. Velleman is clearly guilty of this misreading when he claims that Frankfurt's work is "the best sustained attempt [at] explaining how an agent's causal role supervenes on the causal network of events and states" (Velleman 2000, 132).
[80] Watson 1975.

explicitly concerned with how the capacity for critical reflection and the possibility of alienated agency can fit into the reductive framework of the causal theory (CTA).

To examine Bratman and Velleman's work, let us return again to Velleman's characterization of CTA:

> There is something that the agent wants, and there is an action that he believes conducive to its attainment. His desire for the end, and his belief in the action as a means … jointly cause an intention … which in turn causes the corresponding movements of the agent's body.[81]

Bratman and Velleman argue that, unlike Frankfurt's hierarchical account of the will, CTA provides no way of regulating the agent's motivational psychology in a way that allows only those desires that are a proper part of the agent's self-conception to have causal efficacy. It seems, in fact, that the sole determinant of a particular desire's motivational potential is something inherent in the desire, namely its strength or intensity. This suggests that the most vivid, compelling, or gripping desires are the ones most likely to motivate our behavior at any given time. But such desires are the very ones that threaten the possibility of alienation; and, as Bratman puts it, alienated desires render an agent "less the source of the activity than a locus of forces."[82] So we arrive at the worry that the agent fails to participate in her action because there seems to be nothing other than motivational intensity determining the agent's behavior. This motivational intensity, by its very nature, puts the agent herself on the sidelines of the unfolding of her action. CTA does not tell an adequate story of agency because nothing in the story plays the regulatory role over the agent's desires

[81] Velleman 2000, 123.
[82] Bratman 2001, 312.

that ensures an agent's motivations are truly her own. This is not to say, however, that CTA cannot be fortified to deal with this, in particular by tinkering with the kinds of mental states go into the etiology of action. This is Bratman and Velleman's project.

Bratman and Velleman attempt to improve upon Frankfurt's hierarchical picture by expanding the collection of mental states that tend to be included in the hierarchy as well as in formulations of CTA, i.e., beliefs, desires, intentions. They argue that, while Frankfurt's hierarchical strategy has the right structure, his elementary selection of psychological attitudes prevents his view from serving the purposes of keeping agents in story of their actions. At the same time, they also argue that Frankfurt's view is insufficiently naturalistic because of its reliance on the idea of an agent identifying with her motives.

Recall, the issue is to give an account of an agent actively participating in her behavior (a notion we are still getting a grip on at this point), one that does not make any appeals to irreducible or non-naturalistic phenomena like agent-causation. Frankfurt's hierarchical conception of the will – the idea of an agent identifying with her motivating attitudes via the formation of higher-order desires – can be read as an attempt to pull this off. Bratman and Velleman, however, argue that Frankfurt simply substitutes one notion of agent-causation for another. That is, instead of reducing agent-causation to an event-causal story about mental states, what Frankfurt actually does is "posit self-identification as a primitive mental phenomenon."[83] That is, Frankfurt assumes that when endorsement occurs the agent is necessarily participating; otherwise, the hierarchical view falls victim to Watson's

[83] Velleman 2000, 136.

regress problem – the continuation of the problem of alienation into an infinite string of higher-order desires. As a result of this assumption, they argue, Frankfurt smuggles into the attempt to reduce agent-causation another form of agent-causation in the guise of an agent's endorsing or identifying with her motivations.

Bratman and Velleman extend Frankfurt's view in two ways. First, they add richer, more sophisticated attitudes to the hierarchy. Second, they purge from Frankfurt's picture any appeal to attitudes like identification that presuppose agent-causation. Since, as Velleman notes, "substituting one instance of agent-causation for another, as the target of reduction, does not advance the reductionist project." Another strategy is required. He recommends the following:

> The way to advance the reductionist project is not to substitute one agent-causal phenomenon for another as the target of reduction, but to get the process of reduction going, by breaking agent-causation into its components. And surely, the principal component of agent-causation is the agent himself. Instead of looking for mental events and states to play the role of the agent's identifying with a motive, then, we should look for events and states to play the role of the agent.[84]

So the suggestion is that instead giving an account of something the agent must *do* in order to play an active role in her behavior (this begs the question of agent-causation), what we need to do instead is develop *a reduction of the agent herself* that only appeals to mental states and events. Velleman admits this sounds somewhat bizarre. "Of course, the agent is a whole person, who is not strictly identical with any subset of the mental states and events that occur within him."[85] He argues in response, "a complete person qualifies as an agent by

[84] Velleman 2000, 137.
[85] Velleman 2000, 137-138.

virtue of performing some rather specific functions … [i.e.] the deliberative processes constitutive of agency." He explains:

> the agent's role is to adjudicate conflicts of motives … [therefore] the agent … is that party who is always behind, never in front of, the lens of critical reflection, no matter where in the hierarchy of motives it turns.[86]

We reduce the agent, therefore, by identifying those mental states that can causally interact with other attitudes in such a way that they can be considered "functionally identical" to the agent herself. Thus, we can refer to this kind of view as *representative reductionism*; for it provides a reductive account of an agent actively playing a role in her behavior by appealing to a mental state that can functionally represent the agent in practical thought.

Let us now take a brief look at Bratman and Velleman's specific versions of representative reductionism. Bratman's proposal begins with an assumption of a broadly Lockean conception of personal identity, wherein an agent's temporal persistence is constituted by various connections and continuities among various psychological attitudes. He argues that the way to secure an agent's role in her actions is by adding to CTA's mix certain attitudes that play a particularly important role in the cross-temporal organization of that aspect of an agent's psychology that influences action. Such attitudes, he argues, are higher-order policies to treat certain kinds of desires as reason giving in one's practical deliberation. As he explains,

> this seems a promising strategy in part because such higher-order intentions, plans, and policies have, as a matter of function, tight connections to the temporal

[86] Velleman 2000, 139-140.

> extension of agency. This is why they are candidates for attitudes that, because of their role in our agency, can speak for the agent.[87]

Thus, Bratman focuses on higher-order policies about reasons as the right representatives of the agent for whom they are policies because, given a broadly psychologically reductionist account of personal identity, they are the kind of attitude essentially connected to the agent's personal identity.

Velleman, on the other hand, identifies "the desire to act in accordance with reasons" as the agent's mental representative:

> What really produces the bodily movements that you are said to produce … is that part of you that performs the characteristic functions of agency. That part, I claim, is your desire to act in accordance with reasons.[88]

The characteristic functions of agency he refers to are forms of intervention in the thought process leading to decisions about how to act and in the causal process the leads from decisions to intentions to bodily movements. Velleman argues that the reason he identifies the desire to act for reasons as the attitude that can be the agent's mental proxy in these intervening exercises is that an agent cannot be alienated from this attitude and still be an agent. That is, an essential property of being an agent, on Velleman's view, is having the desire to act in accordance with reasons play a role in the mental causal processes the lead to action.

> [A] person's desire to act in accordance with reasons cannot operate in him without its operation's being constitutive of his agency. What it is for this motive to operate is just this: for potential determinants of behavior to be critically reviewed, to be embraced or rejected, and to be consequently reinforced or suppressed. Whatever

[87] Bratman 2001, 321-322.
[88] Velleman 2000, 141.

> intervenes in these ways between motives and behavior is thereby playing the role of the agent and consequently *is* the agent, functionally speaking.[89]

To put their views in the parlance of PDA, Bratman and Velleman argue that if we restrict CTA's psychological inventory to instrumental beliefs and first-order desires, then agents do not play an active role in their behavior in CTA's story of action. If we add to CTA's mental economy, in a way that does not implicitly assume agent-causation, the theory can overcome this problem. What is needed is a richer moral psychology that recognizes mental states that can conduct "the deliberative processes constitutive of agency."[90] According to Bratman, if we make room for higher-order plans about which desires to treat as reason-giving and allow these plans a role in the causal etiology of an agent's movements, then the agent is active with respect to these movements. Velleman proposes we add to CTA the desire to act for reasons, an *ur*-desire that functions in such a way that it guarantees an agent's active role in her behavior. In both theories, the reason the particular mental states emphasized guarantee the agent's participation in her behavior is their essential connection to the agent's constitution, i.e., to her existing as an agent both at a time and through time. Thus, CTA can overcome PDA and allow the agent to play an active role in her behavior – render her more than "an arena in which certain physical and mental events take place" or " a locus of a series of…causal pushes and pulls" – by adding to its mix of causally efficacious mental states those which have an essential connection to the agent. Having this connection enables these states to *represent* the agent in the reductive causal theory.

---

[89] Velleman 2000, 142.
[90] Velleman 2000, 140.

2.4 The Two Distinctions Revisited

As we saw above, PDA's concept of an agent playing an active role in her behavior can be captured in terms of two distinctions, the active/passive distinction and the robust agency/non-robust agency distinction. Now that we have the essentials of Bratman and Velleman's views on board, we can ask which of these distinctions best captures what they are after. Does the addition of the mental states they recommend correct CTA's inability to account for the basic idea of self-movement – the idea of agent initiating and guiding one's movements – or is their embellishment of CTA intended to capture the more sophisticated phenomenon of robust human action? We can answer this question by returning to case of the unwilling addict, for both Bratman and Velleman take themselves to be responding to this type of case.

I suggest that the proper way of interpreting the case of the unwilling addict is as follows. Despite feeling alienated from her motivating desires, the unwilling addict is still acting intentionally when she takes the drug. This behavior is not wholehearted on her part, and for this reason we can say that an important part of the unwilling addict's character is missing from her action. She plays less of a role in the action because the values that go along with being the person she really wants to be are not making a difference to what she does. But she is not *literally* utterly passive in the face of her desire for the drug. The desire has not rendered her unconscious or hijacked her motor control system. The desire is not a puppet master manipulating her limbs in such a way that she ingests the drugs. Therefore, the unwilling addict's behavior may be *non-robust* because her rational faculties are unable to

get control, but this does not mean it is not *activity*. The fact that her movements fall on the active side of the active/passive distinction means that taking the drugs is something she is doing, not something that is happening to her. If this is right, and if Bratman and Velleman aim to allow CTA to accommodate this kind of case, then it is clear that their concern is with the robust/non-robust issue and not the activity/passivity question.

We can conclude from some of their other remarks that Bratman and Velleman would agree with this. That is, they both seem to think that CTA's inability to capture more than the kind of thing that goes on with the unwilling addict does not make it a failure as a general account of activity or intentional self-movement. Velleman makes it clear that his problem with CTA is not that it precludes the agent from being active *simpliciter*, that the issue is robust agency:

> Of course, every action must be someone's doing and must therefore be such that an agent participates in it, in the sense that he does it. But this conception of agential participation does not require anything that is obviously missing from [CTA]. What is missing from that story is agential participation of a more specific kind, which may indeed be missing from doings that count as cases – albeit defective or borderline cases – of action.[91]

When Velleman complains that "nobody *does* anything" in CTA's story of action, it may sound as if he is claiming that CTA renders agents passive with respect to their bodily movements, but this is not what he means. What he really means to say is that there is a sense of *doing something* that describes cases in which "the distinctly human feature" is present, cases of "human action *par excellence*."[92] From Velleman's point of view, to say that

[91] Velleman 2000, 124.
[92] Velleman 2000, 124.

CTA's guilty of PDA, that it leaves the agent out of the story of action, is to say that CTA cannot capture an agent's full engagement with her rational capacities. In other words, what is missing is not the agent herself *qua* active being; rather what "disappears" is her capacity for robust human action.

Bratman is somewhat less explicit. But if we look at what he says, it is clear that he too thinks that CTA is not guilty of a broader failure to capture activity – that the shortcoming articulated by PDA has to do with the issue of robustness. He writes:

> When a person acts because of what she desires, or intends, or the like we sometimes do not want to say simply that the pro-attitude leads to the action. In some cases we suppose, further, that the *agent* is the source of, determines, directs, governs the action…."[93]

Bratman does not seem to deny that a person can be *acting* when things go as CTA says they do, when "the pro-attitude leads to the action." To say that the agent disappears from CTA's story of action, or to say that she is "merely the locus of a series of happenings" in the story, does not mean that she is literally not being *active* or not exercising her capacity for self-movement. Like Velleman, what this kind of talk is really pointing to is the *robustness* of the kind of agency CTA can accommodate, not whether it can accommodate activity *tout court*.

So Bratman and Velleman agree that even though alienated agency does not qualify as robust human action, it does count as activity. Thus, from their perspective, PDA is not something that afflicts CTA as an account of basic activity; it is only when we want to tell a

[93] Bratman 2001, 311 (emphasis added).

complete story of all forms of human agency, especially those in which our unique higher faculties come into play, that the view's inadequate. Given that this is their concern, I think it is fair to say that they both get a bit carried away when they claim that because CTA does not capture robust agency, agents in CTA's story of action are reduced to "arenas in which events take place" or "loc[i] of a series of happenings". This kind of language is much more apt for raising the charge that CTA does not capture basic self-movement. Agents really would be mere sites where certain events take place in CTA's story of action if the story it tells cannot capture agency is its most basic sense, the capacity to initiate and control self-movement. But this is not the issue they address. That is, their treatment of PDA does not have to do with the active/passive distinction, but with the robust/non-robust distinction.

The fact that Bratman and Velleman raise the robustness issue in terms more suited to question about basic activity and self-movement is a problem. Given that they are two of the most prominent contemporary philosophers of action, one who *is* worried about CTA at the level of basic self-movement might be seduced into thinking they have offered a viable solution to *this* problem. Whether one is satisfied with their solution or not, it now looks as if PDA, understood as a problem about basic activity, is to be solved by tinkering with the details of the mental states upon which CTA draws. But what if the source of PDA's challenge to CTA's ability to capture basic self-movement *just is* the causal theory's exclusive reliance on mental states as the causes of an agent's movements? The prominence of Bratman and Velleman's accounts then seems very unfortunate. For it distracts us from a quite different question about whether CTA captures the way in which agents are active as

opposed to passive when the act. If this is the real issue expressed by PDA, then no matter how sophisticated one makes the motivational psychology of agents in the causal theory, CTA will never capture the agent's participation in her behavior. You cannot get to the penthouse without first making it in at the ground floor.

At this point, let me briefly retrace the steps of this discussion. We saw at the outset that PDA's claim that CTA leaves agents out of the story of action, that it fails to let the agent play an active role in her behavior, is not univocal. I have been focusing on two particular ways of understanding this claim. On one interpretation of PDA, the criticism of CTA has to do with the basic distinction between activity and passivity. The charge is that CTA does not get agents into the story at all because it cannot capture basic self-movement, i.e., the capacity to initiate and control our movements. On the other interpretation, the issue is not about whether agents in CTA's story are acting; rather the issue is about the robustness of the activity agents in this story can achieve. This is to say that agents are left out of CTA's story of action in the sense that the exercise of human agents' higher rational faculties cannot be accommodated by the causal theory. Velleman and Bratman focus on the second interpretation of CTA – the issue of robustness. Unfortunately, however, their rhetoric when giving descriptions of CTA's inability to capture robustness is put in terms that make agents seem completely passive in CTA's story of action. As a result, it is easy to think they offer a solution to the activity/passivity problem with respect to CTA, when what they in fact offer is a solution to the robustness/non-robustness problem. Their views leave open the question of whether CTA is guilty of PDA in the first sense; their views even

distract us from trying to answer this question. In the end, if basic activity poses a problem for CTA, and if the source of the problem is the theory's exclusive focus on mental states as the causes of bodily movements, then Bratman and Velleman cannot help the causal theory.[94]

2.5 CTA & Basic Self-Movement

The purpose of this section is to articulate a way of understanding PDA that remains when we bring to the forefront the distinction between robust rational agency and basic self-movement. I have argued that attempts to deal with PDA have restricted it to the case of robust agency. In the discussion that follows, I bring out a version of PDA that remains an open issue.

---

[94] Before leaving Bratman and Velleman and turning to whether CTA can capture basic self-movement, it is important to note that the failure of their views to attend to this category of activity (the capacity for self-movement) need not be seen as a major shortcoming for their views. The reason for this is that Bratman and Velleman can both respond that basic self-movement simply is not what concerns them, that what they're after is the form of agency that is distinctively human or ideally rational. Considered on their own terms as attempts to give reductive analyses of robust human action, it is possible that Bratman and Velleman's views are perfectly adequate. What matters for my argument is not whether their views, understood as accounts of robust agency, are adequate. What concerns me is the question of whether any account of robust agency suffices as a response to PDA. I am arguing that their views do not suffice. When causal theorists (like Bratman and Velleman) turn their attention to agency by looking into "deliberative processes", they overlook a great deal of what goes on in our lives that deserves to be called agency. This mundane activity essentially involves the exercise of an agent's power of motility or capacity for self-movement, which involves the power to initiate and control her movements. So long as PDA's concept of "playing an active role" in her behavior is exclusively interpreted along the lines of robust human action, the question of whether CTA's able to capture most of what a human agent does goes unanswered. And as I noted above, sophisticated causal theories like those of Bratman and Velleman cannot even get off the ground – the question of defending or rejecting them does not even arise – unless it is shown that basic activity can be understood reductively as bodily movements caused by mental states. So, in the end, my truck with Bratman and Velleman has to do with the level at which their theories are pitched, not with the details of the theories themselves. In fact the strategy of representative reductionism strikes me as the most plausible way of achieving a fully reduced or naturalized account of agency. This is why I pursue a similar strategy later on.

Bratman and Velleman start with reductionist ambitions; therefore, they *take for granted* that our actions are caused by the mental states that motivate us. Velleman's explicit about this: "My objection to [CTA] is not that it mentions mental occurrences in the agent instead of the agent himself … [CTA] cannot be faulted merely for alluding to states and events occurring in the agent's mind."[95] As we have seen, Velleman's confidence here stems from his naturalistic commitments. CTA arises from the need to show that action "consists in some, perhaps complex, causal structure involving events, states, and processes of a sort we might appeal to within a broadly naturalistic psychology."[96] I argued in Chapter 1 that this event-causal analysis of human action is an instance of a broader explanatory principle that characterizes scientific methodology, namely, the idea that the functions of wholes are to be explained by the workings of their parts. CTA embodies this decompositional methodology by accounting for our ability to actively bring things about in the world in terms of subagential entities like mental states (and the neurological mechanisms that realize them).

Notice that CTA's reductionism does not claim that in order for us to do any thinking or acting there must be all kinds of goings on among our parts. There is nothing reductive about the idea that *in order for* whole embodied agents to do things their parts must do things. This is just a claim about the necessary conditions for thought and action; given that we are embodied creatures, it should not be surprising that we cannot do anything without the stuff we are made of doing things as well. The reductive naturalist claim is much

[95] Velleman 2000, 125, 130.
[96] Bratman 2001, 312.

stronger. It says *what it is* for whole embodied agents to do things *just is* for their parts to do things. For many, the kind of reductionism informing CTA is worrisome; even dyed-in-the-wool reductionists should not deny that there is something jarring (at least initially) about the claim that *our* doings are nothing other than causal interactions among *our mental states and bodily limbs*.

The question PDA raises is whether this reductionist view gives any role to the agent herself in the generation of her movement. To begin considering this question, let us start with an example. Mia throws the baseball to the catcher to get the runner out. Mia's throw is an action – something she does – comprising movements she actively performs. Moreover, Mia's throw is an instance of basic activity because she simply noticed the runner heading home and threw. No doubt Mia wanted to get the runner out and thought that by throwing home she could, but she did not need to actually go through this instrumental reasoning in order to act.[97] As an action, Mia's throw should fall within the scope of CTA; the theory should be able to tell a reductive story about what happens when Mia throws to home – a story like this:

[97] It is probably not lost on the reader that I add this sentence as a formulation of a belief-desire explanation of Mia's throw to home. I think it is fair to complain that there is something artificial about the application of this kind of explanation in all actions, and the concomitant insistence that an agent in this situation really can be credited with the distinct mental states upon which the explanation relies. This kind of criticism is starting to bear fruit in artificial intelligence and robotics research. For example, the use of centralized representational states in the effort to design a robot that can competently navigate its environment has proven far less successful than the work by those (e.g. Brooks 1991) whose robots get around by directly responding to their environments. Hubert Dreyfus's discussion of the phenomenon of "everyday coping" in his influential book *What Computers Cannot Do* (Dreyfus 1992) is commonly credited with inspiring this anti-representationalist way of thinking. My defense of CTA is limited to considering PDA, yet a fuller defense of CTA would need to respond to the criticisms from this camp.

(CTA$_{Mia}$) Mia's desire to get the runner out (D(o)) and her belief that throwing home was a way to do this (B(t)) jointly caused the arm movements that constituted her throw to home.

Or,

Mia's throw = D(o)/B(t)-caused arm movements.[98]

With this on board we can now ask PDA's question:

(PDA$_{Mia}$) If D(o)/B(t) caused Mia's arm movements, then what role did *Mia* play in these movements?

As I noted in Chapter 1, CTA's reductionism about agency expresses a commitment to an explanatory and methodological strategy characteristic of modern science, namely, the practice of decompositional analysis. PDA is a specific instance, therefore, of a more general worry about the compatibility of the reductionism of science and the mental and agential phenomena that are the stuff of human existence. The general worry is that the agent doing the acting or the thinker doing the thinking disintegrates into the parts, processes, states, events, etc. that make her up. Nagel nicely voices this worry in an early article:

> If we begin with something that anyone would describe as an action, a man's tying his shoes, for example, we can break it down bit by bit until we have passed beyond the fingers and muscles … to the level of changes in the permeability of cell walls and in the potential gradient at nerve synapses … to alteration in the large molecules at the nucleus of the cell … to the sub-atomic events on which that depends. At some point it will be clear that by traveling deep enough inside the person we have lost him, and are dealing not with the means by which he ties his shoes but with the physiological and mental substructure of his actions.[99]

[98]As I noted earlier, I am assuming that the causal pathway between D(o)/B(t) and the arm movement is "non-deviant".
[99] Nagel 1969, 453.

The question is, at what point do we lose the agent? PDA suggests that we lose her as soon as we try to reduce capacities that we predicate of the whole person to causal interactions among mental states and limbs. For once we go even that far we can ask the kind of question raised by (PDA$_{Mia}$): if Mia's mental states caused her arm movements, then what did *Mia* have to do with them? As one recent commentator puts it, the worry is that

> no causal theory can serve to capture the crucial aspect of agency, the sense in which agents are *active* – as opposed to passive – in the production of their actions… [I]f we analyze [action] as an event caused in some particular way by events and states of an agent, there is always room to ask for a justification for equating those particular events with [movements brought about by] the agent herself.[100]

Importantly, as Nagel's remarks suggest, CTA is not alone here. A similar question can be directed at other endeavors to explain human behavior. For example, we can ask cognitive psychologists and neuroscientists about how movements caused by "motor programs" or "mirror neurons" count as active movements performed by the whole agent.

PDA's question arises because CTA seems to be pointing to two separate and potentially competing causes of a particular bodily movement. It says that the agent's mental states caused the movements, and it allows that the agent herself caused them. Thus, PDA challenges the compatibility of two claims about agency CTA endorses:

1. The movements that constitute intentional actions are caused by mental states.
2. The movements that constitute intentional actions are caused by agents.

We can illustrate this compatibility challenge by drawing an analogy between PDA and Jaegwon Kim's "causal exclusion argument" against the causal efficacy of the mental. Kim

[100] Yaffe 2000, 122.

argues that acceptance of some form of physicalism commits one to mind-body supervenience. He explains this as the view that "what happens in our mental life is wholly dependent on, and determined by, what happens with our bodily processes."[101] The question this raises is, given that every bodily event has a bodily cause, how is a mental cause possible? It appears that mental causes must compete with physical causes, but it is not clear how they can. The mental seems to be crowded out by the physical. Just as it appears that mental competes with (and loses to) the physical if the former supervenes on the latter, so it seems that the agent competes with her mental states if CTA tells the right story about agency.[102]

Now that we have a clear sense of the more basic version of PDA, the question is what CTA can say in response. I develop such a response in the second half of this dissertation, the remaining two chapters.

2.6 Looking Forward

In the next two chapters I offer a way of understanding agents and their parts (including their mental states), that fills in the story those worried about PDA still see

---

[101] Kim 2005, 14.

[102] It is easy for most contemporary philosophers of mind and action to overlook this because the debate about mental causation (wherein Kim's argument finds a home) focuses on the issue of whether we can secure the causal efficacy of mental states *so that* we can ensure our status as agents in the natural world. (For insightful discussions that do not overlook this point see Bayne & Levy 2006 and Horgan Forthcoming.) That is, it is easy to confuse what those who defend the possibility of mental causation within a physicalist metaphysics are after with what the proponent of PDA is after. But PDA argues that even if philosophers succeed in overcoming Kim's challenge to the causal efficacy of the mental (at least within the domain of action), this still will not get us agency, for agency involves bodily movements brought about by *agents*, not caused by the agent's mental states. And, PDA charges, there does not seem to be a story on offer about how an agent's capacity for self-movement can be reduced to the causal powers of mental states.

missing from CTA. This is a story about how it is that movements caused by sub-agential things like mental states can be understood as movements initiated and controlled by the agent herself. The story I offer takes a page from Bratman and Velleman's book insofar as I too aim to tell a reductionist story. A virtue of Velleman's view, in particular, is that it emphasizes the fact that the capacity for agency should not be understood any differently than other capacities we have. Velleman claims that we predicate certain capacities of people, such as the power to digest food and fight off infections, by virtue of the fact that proper parts of us are mechanisms whose function it is to exercise these capacities for us. "When we say that a person digests his dinner or fights an infection," he writes, "we do not mean to deny that these functions actually belong to some of his parts." This is meant to suggest that even though it may seem odd to understand our powers for agency in terms of the capacities of some of our inner mechanisms, this is already the way we understand ascriptions of powers to whole agents. "Similarly, a person may be an initiator of actions – and hence an agent – in the sense that there is an action-initiating system within him."[103]

In what follows I argue that once the Aristotelian conception of the nature of human agents is in place, we can take this talk of an "action-initiating system" in stride. I defend the view that what plays the role of this system in human agents is the faculty of desire. I argue that if we understand the notion of desire in traditional terms that connects to the good, then a case can be made for seeing desires as the agent's reductive representative in the causal process described by CTA. Combining this good-based notion of desire with the

[103] Velleman 2000, 138.

Aristotelian point that an agent's good is a constitutive principle determining the nature and function of its parts results in a picture of desire as the kind of mental state that is essentially connected to the agent's identity. This gives us grounds to ascribe movements caused by desire to the whole agent. Finally, because the view appeals to desire, rather than higher-order attitudes, the resulting account applies to the level of basic self-movement.

THREE

Self-Movement & Constitutive Teleology

3.1 Introduction

At the end of the previous chapter I discussed a version of PDA that applies at the level of basic self-movement. The problem arises because CTA seems to be pointing to two separate and potentially competing causes of a particular bodily movement. It says that the agent's mental states caused the movements, and it allows that the agent herself brought them about. PDA challenges the compatibility of two claims about agency CTA endorses:

1. The movements that constitute intentional actions are caused by mental states.
2. The movements that constitute intentional actions are caused by agents.

Interpreted in light of these two claims the question raised by PDA can be stated as follows: how can we make sense of the claim that the movements performed by the whole are caused by some of its parts while claiming that the whole itself, and not just those causally efficacious parts, is causally responsible for the movements? The answer, which I formulate in this and the next chapter, is that some parts are related to the essential nature of the whole in such a way that what they do counts as the whole itself doing it. The purpose

of this chapter is to lay out the conceptual groundwork for this response to PDA. At the center of the discussion is the claim that self-movement is a normative concept.

Causal theorists, of course, reject the notion that there is any competition between agents and their mental states in its story of what happens when someone acts. In this chapter I argue that they can defend their view by appeal to the fact that CTA is an instance of a widely accepted picture of behavior, one that accounts for the general capacity for something to move on its own. I'll refer to this view as *the standard story of self-movement*. If CTA can align itself with the standard story, this puts proponents of PDA in a tough position; for it would commit them to the claim that what happens when human beings (and any other agents within CTA's explanatory scope) move themselves is a phenomenon that cannot be accounted for by a theoretical model that seems to account for the behavior of every other kind of self-mover.[104] As an instance of the standard story, CTA shows human agency as continuous with the way the rest of the world works. That is a strong reason for preferring it to the kind of non-reductive view suggested by those who raise PDA.

In this chapter I present a normative (Aristotelian) interpretation of the standard story of self-movement. I aim to garner support for this interpretation of the standard story by demonstrating how it allows for a unified account of the self-movement of three different types of agents: artifactual agents, conventional/collective agents, and organic agents. The particular focus in this chapter is on artifactual agents like like robots and conventional/

[104] This is not to say that by aligning itself with the standard story of self-movement CTA must be committed to the view that the explanation of human behavior is no different from that of, say, an autonomous robot or a honey bee. Surely there can be more and less sophisticated versions of a single form of explanation.

collective agents like teams and corporations. I leave treatment of organic agents – the class into which human agents obviously fall – for the following chapter. The reason I start by focusing on self-moving artifacts and collectives is that it is easier to make explicit an essential feature of my normative interpretation of the standard story, one that shows up in the part-whole relation that exists between self-moving agents and their parts, when considering these kinds of agents than when the focus is on organisms. This is the idea that self-moving agents are *teleologically constituted* – the nature and function of parts of a self-mover are essentially related to the overall ends and aims of the agent as a whole. Therefore, *constitutive teleology* characterizes the conditions that must be in place for us to apply the standard story of self-movement, that is, for us to be able to say, of some particular internal cause, that the movements it causes are properly counted as movements the whole agent is performing. It is this idea of constitutive teleology that is more easily made conspicuous in the case of artifacts and collectives than it is in the case of organic agents.

3.2 The Standard Story of Self-Movement

At the heart of my analysis of the standard story is the idea that self-movement is a normative concept, i.e. the determination of whether or not something's movements are correctly identifiable as genuine self-movement depends on certain *normative* considerations. The purpose of this section is to articulate the normative elements of self-movement and identify the conditions upon which these normative elements depend. Along the way, I

examine, in this and the remaining sections, how these conditions are specifically instantiated in two different kinds of self-moving agent, artifacts (robots) and collectives (teams).

It is important to note that it is no part of my argument that my Aristotelian interpretation of the standard story is the *only* way to account for what happens when agents purposively move themselves; or, that there are not cases of self-movement that seem to work differently than how the standard story says self-movement works. It is possible that there are cases of self-movement that do not involve any of the normative considerations I go on to discuss. I do think, however, that *paradigmatic* cases of self-movement, such as what is on display when autonomous robots or the Mars Rover or cheetahs or police officers navigate their environments, do in fact involve the normative framework I develop. There is a common structure to these core kinds of cases that is revealed by the normative analysis I present.

Dretske gives a clear statement of the standard story of self-movement: "behavior is endogenously caused movement, movement that has its causal origin *within* the system whose parts are moving."[105] This seems obviously right. Just compare, for example, the difference between intentionally taking a seat on the lawn and being pushed to the ground by a strong wind. It is clear that "the locus, internal or external, of the cause" of movement is the salient difference between the two cases – one a case of voluntary behavior, the other a case involuntary movement. [106] Thus, we have the standard story's basic thesis:

[105] Dretske 1998, 2.
[106] Dretske 1998, 10.

(BT) Self-movement is internally caused bodily movement.

I refer to this as the *basic* thesis to highlight the fact that, in order to arrive at a complete account of self-movement, we need to go beyond the claim that a given bodily movement counts as self-movement if it is the effect of a cause internal to the mover.

We can begin adding to (BT) by attending to the difficulties that go along with trying to make sense of the notions of "internal" and "external" at work in the standard story. What exactly does it mean to say that the cause of movements comprising self-movement is internal to the agent? Dretske notes, "*internal* [cannot] simply mean inside or underneath the skin, fur, fins, feathers, or whatever" of an agent.[107] Too see this imagine that Scooter, my Jack Russell Terrier, swallows an electrode that can send signals to his motor cortex resulting in the wagging of his tail. Here we have endogenously produced movement, movement with an internal cause, yet we do not want to say that the movement is something Scooter does or performs. What needs to be added to the picture is the idea that the internal cause must be "a proper or integral *part* of the system exhibiting the behavior."[108] The problem with the electrode is that it is not connected to Scooter in the right way for it to be the right kind of thing to cause self-movement. Specifying that the internal cause must be a proper *part* of the mover, and not just something attached to or inside of it, gives us our first addition to (BT):

(BT+P) Self-movement is movement caused in the right way by a proper part of the mover.

---

[107] Dretske 1988, 3.
[108] Dretske 1988, 3.

There are reasons to think that, like (BT), (BT+P) also fails to specify a sufficient condition for self-movement. To see this, consider another example, a case of a self-moving artifact. We do not say that just any movement a robot, machine, or other artifact makes counts as self-movement (assuming it is the kind of thing some of whose movements we think do have this status). When a mobile robot's in good working order, it seems reasonable to say that when the mechanisms responsible for its locomotive capacities are performing their functions, the robot is moving itself. Now let us imagine a robot that is not in good working order, one that has a fault in the mechanism responsible for preventing the robot from overheating. We can imagine that this mechanism malfunctions in such a way that it disengages the motor-control mechanism and takes over the function of causing the robot to move about. We can even imagine that, at least in the short term, these randomly generated movements somehow manage to mimic the very same movements the robot would make if its movements had their usual causal origin. Here we have a case of internally caused movement by a proper part, but I do not think we should say this counts as a real case of the robot's self-movement (I will explain why in a moment). Comparing these two cases suggests that the way to further refine the basic thesis (BT) and improve upon (BT+P) is to add requirement (R): self-movement must be caused by the *right* part. This gives us the following:

(SM) Self-movement is movement caused by the right proper part of the mover.[109]

[109] As I noted earlier, I am assuming that that the causal chain is "non-deviant". Thus, the reader can attach the phrase "in the right way" whenever I speak of a mechanism causing movement.

To recap: we began with the basic thesis (BT) at the heart of the standard story, the claim that self-movement consists in bodily movement generated by an internal cause. The case of Scooter swallowing the electrode illustrated why (BT) is not sufficient. What is "internal" about the cause of movements that count as self-movement goes beyond being literally under the agent's skin. It also involves being connected to the agent in a certain way. Thus, we must add to (BT) the requirement (P), which states that the internal cause be a proper part of the agent. The example of the robot's malfunctioning thermostat demonstrated why even (BT+P) does not adequately account for self-movement. The lesson of the robot example is that movements that are ascribable to the whole agent, movements that qualify as self-movement, are not brought about by just any part of the agent. We need to add to (BT+P) the requirement (R): movements must be brought about by a particular part of the agent – the *right* part.

3.3 The Proper Cause of Self-Movement

At this point we have a general picture of the standard story of self-movement. As a resource from which to generate CTA's response to PDA, however, this general picture is inadequate. Without a further story about what it means to say that mental states are the right internal causes, or about the criteria for qualifying as the proper kind of internal cause, (SM) does not tell us very much. As I mentioned at the outset of this chapter, we can interpret PDA as posing the following question to CTA: Why do mental states not crowd out or compete with the agent in CTA's story of what happens when someone acts? Saying

that mental states are the right internal causes does not provide a substantive answer to this question. So we are left with an important question:

(Q_1) What makes a mechanism the right cause of self-movement?

Answering this question will involve making explicit the normative underpinnings of the concept of self-movement I mentioned earlier.

To begin with, it is important to note that what makes a certain mechanism the proper source of a robot's self-movement is not a matter of stipulation; it is not a matter of an arbitrary assignment of that status to one of its parts. If this were the case, then there would be no real difference between the movements caused by a malfunctioning thermostat and those brought about by a robot's motor-control mechanism. There would not be anything about the movements themselves or their causes that would distinguish the whole agent actively moving itself from the whole agent passively being moved by one of its parts. But there are differences here. For one thing, beyond being a cause of the robot's movements, being the right movement-causing mechanism involves being hooked up to other mechanisms that perform functions that assist with or benefit from the resulting movements. This is important because a mobile robot lacking connections between, say, its motor-control and its perceptual systems will not be able to reliably and effectively serve any of the purposes or perform any of the functions for which it was designed.

In the robot example I have been considering, the faulty thermostat manages to cause movements of the robot that mimic the movements the motor-control mechanism would cause given the robot's environmental conditions. Unlike the motor-control

mechanism, however, the faulty thermostat is not connected – at least in the same direct way – to the robot's perceptual or sensory systems. This is why we can say that the fact that the movements caused by the faulty thermostat manage to get the robot around its environment successfully is an accident. For the movements' characteristics, say, their speed and direction, are not a result of the robot appropriately reacting to its surroundings. As a result, there is nothing purposive or deliberate about the movements.[110] This is evident when we note that artifactual agents like mobile robots are designed to perform certain functions, to do certain jobs, to pursue certain ends. This is why they are built with functional parts designed to support capacities like perception and self-movement. Thus another way of articulating the difference between the movements caused by the faulty thermostat and those caused by the motor-control mechanism is to say the first set of movements do not contribute to the performance of the functions the robot's designed to perform – at least, that is, they do not do this in any reliable way. Though these random movements happen to mimic what the robot would be doing if the motor-control mechanism was in play at the moment in these particular conditions, it is obvious that – barring something close to a miracle – this congruence will not last for long. As a result, the robot will not do what it is designed to do.

So the idea of the proper or the right cause of self-movement mentioned in (SM) and (Q_1) can be made out by appealing to the overall functions a robot is designed to perform.

[110] For those uncomfortable applying concepts like 'purposive' or 'deliberate' to the movements of a robot can substitute them with concepts like 'sub-purposive' or 'sub-deliberate'. This reserves the full-blown version of the concepts for the case of human or rational agency, while respecting the thought that some sophisticated self-moving robots can react to their environments.

These functions are essentially related to the nature and function of (at least some of) the robot's parts, including, of course, the part responsible for causing its movements. This implies that the proper cause of self-movement is the mechanism whose proper function is to generate movements that, in a reliable and non-accidental way, constitute or contribute to the performance of the robot's own proper functions. Characteristic of the mechanism with this proper function is being appropriately connected to perceptual and other systems within the robot that allow it to respond to the layout of its environment. Having this proper function and these systematic connections are the marks of the right internal cause of self-movement.

Thus, in response to Q_1 we can give the following answer:

(A_1) The proper cause of self-movement – movement caused by the whole agent – is the mechanism whose proper function is to generate movements that characteristically constitute or contribute to the proper function(s) of the agent as a whole.

3.4 Bodily Movements vs. Behavior

Before making some further refinements to the standard story of self-movement, I need to deal with a confusion engendered by (A_1)'s talk of "movements that constitute or contribute to the proper function(s) of the agent as a whole." This will require introducing the concept of *behavior*.

(A1) tells us what it is about a particular element within a self-moving agent's functional architecture that qualifies it as the proper cause of self-movement. The answer is problematic, however, because the property of causing movement that contributes to (supports, constitutes, etc.) an agent's proper functioning is not unique to those parts of agents – like the motor control mechanisms in mobile robots – that generate movements that would intuitively strike us as movements the agent actively performs. For example, the proper functioning of both a robot's motor control mechanism *and* its thermostat generate movements that support the proper functioning of the robot as a whole. To see this, let us say that the thermostat performs its proper function of regulating the robot's internal temperature by switching internal fans on and off and by moving varying amounts of coolant through a radiator. According to the criteria articulated in (A1), spinning the fan and moving the coolant count as self-movement, active performances attributable to the whole robot and not just to some of its parts. I think it is safe to say, though, that these internal goings-on caused by some of a mobile robot's sub-systems are not what we have in mind when we say a mobile robot is capable of self-movement. What we do have in mind are activities like grasping objects and navigating the environment, processes that we easily recognize as the kind of things whole robots do.[111]

[111] This is not to say that there could not be a robot for which temperature regulation is within its capacity for self-movement. This would involve the robot's having the ability to monitor and process information about its temperature – in the way it can monitor and process information about its environment – and to switch on its internal fan in response to this information – in the way that it moves its limbs in response to information about its environment.

The problem with (A1) is perhaps even more striking when it comes to animal agents like cheetahs or human beings. We can see this if we allow ourselves to extend talk of proper functions to living things and their parts (an extension I develop more fully in Chapter 4). Following Aristotle, we can say that a living thing functions properly by pursuing its good, a state of flourishing that involves the full exercise of the capacities and powers characteristic of a creature of its kind. Moreover, on this view the proper function of every proper part of an organism contributes to its good. Borrowing a concept I will discuss in Chapter 4, I will refer to the proper cause of self-movement in living agents as *the faculty of desire*, a sub-system in animals akin to the motor control mechanism in robots. According to (A1), movements caused by the faculty of desire count as movements performed by the animal as a whole because the proper function of this faculty is to generate movements that contribute to the animal's good. If this is true, though, then *every* organic part of an animal is a proper cause of self-movement. For example, just like a properly functioning faculty of desire, a properly functioning digestive system generates movements that contribute to the good of a living agent. These movements include events like the contraction of muscles along the digestive tract and the release of digestive fluids within the stomach. (A1) implies that these movements of an animal's internal mechanisms count as self-movement in the same way that the movements of our arms and legs caused by the faculty of desire when we walk across the room do. That is, in terms of the agent's involvement with the movements, the agent plays just as active a role in digestion as she

does in intentional action. Because this example so vividly makes the point, I will refer to this problem of (A_1)'s over inclusiveness as *the digestion problem*.

In order to avoid the digestion problem, we need to constrain the concept of bodily movement in a way that allows us to hold on to the proper function criterion at the heart of (A_1) while limiting the kinds of movements to which this criterion applies in a way that avoids being *ad hoc*. I suggest we can do this by appealing to the concept of *behavior*. The idea is that only those bodily movements that first qualify as behavior can also count as self-movement. Behavior is an elusive concept, and there is no general consensus among behavioral scientists and philosophers over the types of animal movements that qualify. In light of this, I will not attempt to provide a full analysis of the concept of behavior that gives its necessary and sufficient conditions.

We can begin by returning to Chapter 2. There I articulated a spectrum of bodily movements that went from robust human action, through basic self-movement, to involuntary movement. The point of this spectrum was to elucidate the notion of "agential involvement" at the heart of PDA, i.e., what it means to say that an agent plays a more or less active role in her movements. We can begin to understand *behavior* by noting that the term does not denote a form of movement that shows up at a distinct point along this spectrum. That is, it is not a category of bodily movements whose criterion is the extent to which an agent is actively involved with the initiation or control of her movements. Rather, it is a broader category that subsumes all voluntary movements and some involuntary movements as well.

One way to grasp behavior is to consider the concept of an *ethogram*, a diagnostic tool used by ethologists when studying animals in their natural habitats. An ethogram is generally defined as a catalogue of discrete movements performed by a certain kind of animal, a catalogue whose entries are the result of observation of the animal going about its everyday life in its environment. There is debate among ethologists and experimental psychologists over how the movements catalogued in an ethogram should be described. Some argue that anything beyond the barest physical description involves making unwarranted assumptions about the function of the movements being observed, while others argue that invoking functional concepts like "grooming", "mating", "eating", and "fleeing" is in fact justified after sufficient observation. Most agree, however, on an ostensive definition of behavior, such as the following:

> Behavior =$_{df}$ observable, molar-level movements that go on when an animal performs its characteristic activities; normally consists of movements responsive to environmental information an animal receives via its perceptual systems.[112]

I take no stand here on the debate about the degree to which, for the purposes of the science of animal behavior, we must invoke functional categories when describing these movements. It is important to see, however, that an animal need not be performing a functionally specifiable vital activity – eating, seeking shelter, mating, vocalizing, hunting,

[112] Though he does not use the word 'behavior' to describe it, Aristotle would accept this definition as an apt characterization of the type of movement treated in his *De Motu Animalium*. Support for this comes from the fact that Aristotle does not think we find such movement in the hierarchy of living things until we move from the nutritive soul shared by plants and animals to the sensitive soul endemic to living things with perceptual capacities.

foraging, etc. –in order for its movements to qualify as behavior. The reason for this is that, just like most of us humans, self-moving animals may sometimes simply move for no reason at all. Just as I may mindlessly tap my finger on my desk, my cat Sam may flick his tail simply because he can. Similarly, an ape may grunt, a horse whinny, or a cheetah growl without trying to communicate in any way. Of course, idle movements like this may involve making the same movements an animal would if it was performing a vital activity. However, I see no reason to deny that, though these movements serve no vital function in this case, they still qualify as behavior.[113]

So behavior consists of those observable bodily movements an animal performs when it goes about its everyday activities, whether or not, on any given occasion, the movements serve a purpose like attaining nourishment or fleeing a predator. In addition to not having to be functional, movements do not need to be voluntarily brought about or actively performed by the animal when they are occurring in order to count as behavior. Dennett, for example, considers the routine involved in the egg laying behavior of the digger wasp (*Sphex ichneunmoneus*).

> When the time comes for egg laying, the *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away … [T]he wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If the cricket is moved a few inches away while the wasp is inside making her preliminary inspection, the wasp, on emerging from the burrow, will bring the

[113] I suspect there is a strong evolutionary argument in support of the idea that only animals capable of functionally specifiable movements that constitute the performance of vital activities can perform mindless idle movements.

> cricket back to the threshold, but not inside, and will then repeat the prepatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again she will move the cricket up to the threshold and re-enter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion this procedure was repeated forty times, always with the same result.[114]

This kind of routine is what ethologists call a *fixed action pattern*, a series of movements that an animal either learns to perform in certain situations, or, more often, is a part of the animal's neurological hardwiring. Other examples include things like mating dances that males of various species perform in order to attract mates. It is clear from the wasp example that, perhaps to varying degrees, there can be sophisticated movements going on without the animal being in complete control of them. Once the egg laying routine gets started, or once the step where the wasp checks the burrow occurs, the wasp is programmed to move in certain ways. This is what I am calling attention to by saying that such movements are involuntary – there is no sense, even in whatever attenuated sense is appropriate to the kind of animal at hand, in which the animal has complete control over the initiation and guidance of its movements. At the same time, the movements fit the criterion for behavior. They are the kind of movements that go on when an animal is going about its normal active interaction with its environment.

My discussion of behavior has focused on the way in which the concept is used by those who study animals in their natural habitat. I think this ethological conception of behavior is most intuitive and most familiar, insofar as it is situated in a practice and a point

---

[114] Dennett 1984, 11.

of view familiar to us all, namely, observing how animals live while immersed in their environment. Of course, the types of animals and environments most of us are exposed to are much less exotic than those experienced by practicing animal psychologists. But the selfsame concept of behavior used to chart the comings and goings of lions on the Serengeti applies to the daily lives of squirrels in Central Park. Notice, however, that this ethological conception of behavior can also be extended to cover the activities of artifactual agents, such as the sophisticated mobile robot I have mentioned. The kind of robot I have been considering is designed to perform certain functions that involve manipulating and navigating through the environment in which it is situated. This is why the designers of these robots equip them with sensors that allow their motor control mechanisms to coordinate the robot's movements with the lay of the land. Given that this kind of robot has some kind of commerce with its environment, and has the capacity to move in ways that are responsive to this environment, its movements also meet the ethological criteria.[115]

We can now return to (A1). As I mentioned at the beginning of this section, (A1) is problematic as a statement of what it is to be a proper cause of self-movement because its proper function criterion is over inclusive. There are many types of bodily movements that contribute to the overall ends of an agent that we would not characterize as self-movement. The clearest examples of such movements are those like the contractions of muscles in the digestive tract or the regular and rhythmic beating of the heart caused by the autonomic

[115] Returning to a point I made above, movements that are not responsive to the environment via sensors, such as a robot's internal fan turning on every thirty seconds, do not meet the ethological criteria and, therefore, do not count as behavior. I owe this point to Wayne Davis.

nervous system. These movements happen automatically without any participation on the agent's part. (This is why people can survive in a permanently vegetative state for years.) I have appealed to the concept of *behavior* in order to restrict the class of movements to which (A_1) applies. Thus, we can now move from (A_1) to (A_1)*:

(A_1)* The proper cause of self-movement – movement caused by the whole agent – is the mechanism whose proper function is to generate behavior that characteristically constitutes or contributes to the proper function(s) of the agent as a whole.

Thus, the response to the digestion problem is that the movements that constitute sub-agential processes like digestion – basically those caused by the autonomic nervous system – are not behavior. Since they are not behavior, they are not the kind of movements that can qualify as self-movement.

I will return to the digestion problem again at the end of Chapter 4, where I consider it in light of the Aristotelian account of self-movement I develop in response to PDA. For now we can move on to question of how behaviors caused by the proper cause of self-movement, given that they are still caused by a mechanism within the agent, count as movements caused by or performed by the agent herself.

3.5 Why Some Behavior Counts as Self-Movement

The purpose of this chapter is to put in place a conceptual framework that will allow us to respond to PDA at the level of basic activity, i.e. the problem left standing at the end of

Chapter 2. I have argued that at the heart of this framework is (SM), the standard story of self-movement. The standard story's basic thesis (BM) is that self-movement consists of internally caused bodily movements. I noted that there are problems with this basic thesis, and sections 3.3 and 3.4 have been dedicated to refining this basic thesis by giving an account of the proper cause of bodily movements that count as self-movement. This discussion has yielded (A_1)*. There is still another step to take, however, in order to respond to PDA. The reason for this is that PDA's skepticism about CTA's ability to capture basic activity is not directed at CTA's appeal to mental states *as opposed to some other type of internal cause*. If this were the issue, then (A_1)* would suffice. That is, (A_1)* responds to a worry of someone already convinced of the basic thesis (BT), someone who is comfortable with the idea of internal causes of self-movement in general. But a proponent of PDA is not comfortable with this. Therefore, a full response to PDA must address a further question:

(Q_2) Why does behavior caused by the mechanism in (A_1)* count as self-movement caused by the whole agent?

(Q_2) asks for an explanation of how we go from the idea of a mechanism whose proper function is to cause behavior with a certain relation to the proper functioning of the whole agent to the conclusion that movements caused by such a mechanism count as movements brought about by the agent as a whole. Like (A_1)*, the answer to (Q_2) appeals to the idea of agents and their parts having proper functions.

We can start on an answer to (Q_2) by making explicit the broadly Aristotelian view of the nature of objects that is in the background of (A_1)*. According to Aristotle, there is an

essential connection between what something is and how it functions: "the nature of a thing is its end or purpose" (*Physics* 2.2 194a). That is, an object is the kind of thing it is – it has a particular nature or identity – by virtue of what it does, i.e. its purpose, function, or characteristic activity. This is most evident in the case of artifacts. Being a house or a thermostat or an automatic transmission is not simply a matter of being composed of certain materials or being a certain size and shape; rather, it is a matter of being the kind of thing that performs certain functions for which it is designed – provide shelter, regulate ambient temperature, control the gears in a drive-train. In order to serve its purpose, an object must comprise an organized structure of functioning parts, and having this functional architecture renders an object a unified whole as opposed to, as Aristotle puts it, a mere heap. So we have two relations here: one between an object's function and its nature; the other between an object's function and the organized, unified structure of its working parts.

With this brief sketch of the Aristotelian background to $(A_1)^*$, we can return to (Q_2). $(A_1)^*$ relies on a certain relationship between the mechanism responsible for self-movement and the overall ends and aims of the self-mover of which it is a part. The proper function of the mechanism in $(A_1)^*$, and, therefore, the material structure enabling the performance of its function, can be explained by the overall proper function of the agent. Thus, the nature and function of (some of) a self-mover's parts – what they do and how they are structured – depends on the overall aims and functions of the self-mover as a whole. We can refer to this part-whole relation as *teleological constitution* because the self-mover's *constitution* – the nature of its parts and the way they are structured – is essentially related to its *telos* – its proper aims

and functions. This relation of constitutive teleology between an agent's behavior generating mechanism and its purposes and functions is essential to what makes it appropriate to identify a particular mechanism as the proper cause of the agent's self-movement.

The concept of teleological constitution allows us to explain how parts caused by the mechanism responsible for self-movement redound to the agent as whole. The explanation runs as follows. The proper function of an agent consists in the pursuit of certain aims, the achievement of certain ends, and the performance of certain activities characteristic of its kind. As my sketch of the Aristotelian view illustrated, these aims, ends, and activities that characterize the proper functioning of a given agent are intimately connected to the agent's nature and identity. Part of what it is to be an agent of a certain type is to be the kind of thing whose proper function it is to perform certain activities and pursue certain aims. Agents with proper functions are teleologically constituted, meaning they are made up of parts whose nature and proper function are essentially related to the agent's proper functioning. Teleologically constituted self-movers contain a part or mechanism whose proper function is to generate behavior that either constitutes or plays an instrumental role in the agent's proper functioning. This is the mechanism, the proper internal cause, described in $(A_1)^*$.

So the mechanism described in $(A_1)^*$ has its own proper function, one that is related to the proper functioning of the agent of which it is a part. Given the relation between an agent's nature and its proper function, we can say that the essential nature of the mechanism described in $(A_1)^*$ is intimately tied to the nature of the whole agent. The agent's nature

involves pursuing certain ends and partaking in certain activities, and the mechanism's nature involves generating behavior that puts the agent in pursuit of those ends or that constitute participation in these activities. As a result of this *congruence between the natures of the proper behavior generating mechanism and that of the whole agent*, we can say that movements caused by this mechanism (when it is properly functioning) redound to the whole agent as cause because they are generated by something that essentially embodies what makes the agent the kind of thing it is. That is, behavior caused by the proper mechanism are movements caused by the agent's nature; and in that sense they are movements ascribable to the agent as a whole. In other words, movements caused by a mechanism whose own nature is essentially tied to the agent's nature count as movements caused by the agent itself.[116]

We can examine constitutive teleology further by considering an example. Imagine a self-moving robot whose job is to load and unload baggage from the hulls of commercial airplanes. Obviously, in order to do this job, the robot has to be able to move about, to go from, say, the airplane to the baggage receiving dock at the terminal. In order to do this with any success, however, the robot must have some ability to sense its environment so that it can competently navigate around the many obstacles in its path and stay out of the way of incoming and outgoing aircraft. Given its overall proper function – to move baggage from planes to the terminal – the robot must be equipped with mechanical parts that have their own proper functions. For example, our robotic baggage handler must have certain environmental sensors and a motor-control mechanism that receives input about the nature

[116] I argue in Chapter 4 that this statement has to be clarified when we turn to human beings.

of the robot's immediate surroundings from these sensors. Providing this data about the lay of the land is the proper function of the robot's sensors, and generating movements that respond appropriately to this data is the proper function of its motor-control mechanism. We can say, then, that the robot's proper function of transporting baggage from planes to the terminal is its *constitutive principle*. That is, it is a principle by which we can understand the many functional parts that together make up the robot – understand why it has the parts that it does, and why these parts have the proper functions that they do. The nature and function of the parts is essentially related to the proper function of the whole agent of which they are parts. This is the heart of constitutive teleology.

Notice that the introduction of the notion of constitutive teleology into the discussion brings us to the *normativity* of the concept of self-movement. What is normative about the concept of self-movement is that the application of the concept to a given bit of movement presupposes a normative framework inherent in the self-moving agent – constitutive teleology. This means that the concept of self-movement (at least in its paradigmatic applications) presupposes entities with *ends, purposes, and proper functions*. Thus, in addition to an appeal to an internal source of movement, we also need to be able to situate movements within a framework that allows us to identify certain characteristic functions or activities that the kind of thing doing the moving is supposed to do. Identifying a particular mechanism as the *right* inner cause or the proper source of self-movement involves showing that the proper function of the mechanism is to generate movements that serve the overall ends and purposes of the agent.

(Q_2) asks why the behavior caused by the proper mechanism count as movements brought about by the whole agent herself. The relation of constitutive teleology that characterizes self-moving agents allows us to give this answer:

(A_2) Behavior caused by the mechanism described in (A_1)* counts as movements caused by the whole agent (self-movement) because they are movements caused by the agent's nature.

Let us return for a moment to PDA. Proponents of PDA maintain that self-movement is not caused by internal states, events, or processes; rather it is caused *by the agent herself*. This is what makes it *self*-movement. The standard story rejects this picture because it understands movements caused *by the agent herself* as a subset of movements *caused by internal mechanisms, states, events, or processes*. What is distinctive about self-movement – what makes it the case that it is ascribable to the whole agent – is the fact that the movement plays a role in the overall life of the agent by enabling the agent to perform its proper function. So according to the standard story, what makes a bit of behavior *self*-movement is not that it is caused by something irreducible, something identifiable as a "self", but that the movements are intimately connected to the agent's proper function, the ends and aims it is designed to achieve. What makes some internal mechanism the proper cause of movement that counts as self-movement is that when *it* properly functions the agent behaves in ways that contribute to, enable, or constitute the performance of the *whole agent's* proper function(s). The agent's proper function is intimately connected to its nature – what it is to be an agent of a particular kind is to be the kind of thing with a particular proper function.

3.6 Constitutive Teleology & Collective Agents

In this section I introduce a social practice-based account of collective agency in order to further illuminate the normative concept of constitutive teleology at the heart of the standard story of self-movement. I focus on the following question: what enables a single individual agent to act in such a way that her so acting constitutes a doing by the entire collective of which she is a member? The answer to this question will help us better understand (A2).

We think that collective entities like teams, governments, and corporations can do things – score goals, fight wars, bring rivals to court – when their parts, i.e. the individuals who are their members, do something. As Rob Wilson explains:

> Banks can foreclose on your mortgage, city councils can raise your property taxes, and Her Majesty's Government can request the pleasure of your company. As economic, political, and legal entities, each of these agents can bring about effects, sometimes effects that matter a great deal to us. They act through the agency of individual persons, to be sure, but it is only as a representative of a bank, a council, or Her Majesty's government, that the acts of particular persons count as foreclosing our mortgage, raising our property taxes, or imprisoning us.[117]

The account of collective agents that I lay out is not meant to be an analysis that captures collective agency in all its forms. I limit my discussion to collective agents that arise within the context of explicitly rule-governed *practices* or *conventions* like games. We can refer to them as *formal* collective agents. Typical examples of formal collective agents are teams, universities, governments, and other institutions. What they all have in common as

[117] Wilson 2005, 60.

collective agents is that they are groups comprising individuals whose coordinated behavior constitutes participating in activities, occupying certain roles or offices, and performing certain functions that would not exist but for a given set of rules or norms. Therefore, what I have to say about collective agents will not necessarily extend to, say, a group of people rioting. There are certain activities that can come about, and, likewise, certain goals that can be achieved, when people are rioting that cannot arise from a single rowdy individual. So when it comes of rioting, talk of collective agency seems appropriate. But a group of individuals rioting (or looting), is an *informal* collective agent because it's existence is not due to conformity with explicit rules or norms.

Social practice based accounts of collective agency are an alternative to individualist accounts of collective agency. An important difference between the two forms of account is the role that the intentions of individual agents play in making collective agency possible. According to individualist accounts, a sufficient condition for collective agency is the presence in each participating agent's psychology of an intention whose content is roughly "I intend that we Ø" or "I intend to do my part in our Ø-ing."[118] On the individualist view, collective agency is a bottom-up affair, which is to say that certain conditions of the individuals that comprise the collective must be in place prior to the emergence of the collective. A social practice account denies the individualist claim about the necessary presence of certain intentions of participating agents. This is not to say that the account deems such intentions irrelevant. However, unlike the individualist account, the social

[118] See Bratman 1999.

practice account allows collective agency to arise from a group of individuals participating in established rule-governed practices, ones in which individual agents play certain practice-dependent roles or perform certain functions. The animating idea behind the social practice account is that whether an individual is playing such a role, at least at a particular time, depends on the environment or context in which the individual is situated. For example, in certain conditions a hockey player can make a shot she was actually trying to miss, thereby unintentionally making it the case that the hockey team does something, namely, wins the game. People do not accidentally get in this kind of situation, so at some point along the line prior to this particular case of collective action, the individual player's intentions were important. Nonetheless, in this example there is collective agency without an agent having the kind of intention essential for the individualist account.[119]

The concept of constitutive teleology can be made particularly conspicuous by examining formal collective agents because, unlike robots, formal collective agents completely owe their unity as agents to norms. That is, formal collective agents do not have the kind of physical boundedness or discreteness we see in robots and other artifactual agents, and, therefore, the existence of this type of collective agent as a unified collection of functional parts emerges solely from the ways in which these parts are normatively bounded.[120] This notion of normative boundedness lies at the heart of constitutive teleology. Constitutive teleology binds the movements, functions, and processes, i.e., the general goings-on within an entity into the animated life of a self-moving agent.

[119] See Stoutland 1997 and Haugeland 1998.
[120] This is true of what I call below *formal* collective agents, but not necessarily true of *non-formal* collectives.

We can get our initial bearings on formal, practice-based collective agency more generally from the influential discussion of practices in the early work of John Rawls and John Searle. They explicate the idea of a practice in terms of rules. Rawls, for example, defines the concept of a practice as follows:

> I use the word "practice" throughout as a sort of technical term meaning any form of activity specified by a system of rules which defines offices, roles, moves, penalties, defenses, and so on, and which gives the activity its structure.[121]

In order to understand the rules at the heart of a practice, it is important to recognize a distinction Searle draws between *regulative* and *constitutive* rules.[122] The difference between these two kinds of rules (or norms) has to do with the way in which they are related to the activities to which they are applied. Regulative rules apply to antecedently existing forms of behavior. People were walking, for example, long before there were rules like "do not walk on the grass" or "only pass on the left". Constitutive rules, on the other hand, are "logically prior" to particular cases of behavior falling under them. This is because such rules *create* or *make possible* the activities to which they apply; they "define new forms of behavior" (Searle) and "specify…new form[s] of activity" (Rawls). So, for example, prior to the creation of the game of baseball people could walk from one spot to another; they could not, however, *advance to first base*. The rules of baseball bestow (what Searle calls) a *status function* that applies to certain instances of walking from one spot to another. As Searle explains, a status function is

---

[121] Rawls 1999, 20.
[122] Searle 1995, 27-29.

> a special kind of assignment of function where the object or person [or activity] to whom the function is assigned…can perform the function only in virtue of the fact that there is a collective assignment[123] of a certain *status*, and the object or person performs its function only in virtue of collective acceptance by the community that the object or person has the requisite status. These assignments typically take the form *X counts as Y.*[124]

We can explicate this idea with a spatial metaphor. The constitutive rules of a practice institute a kind of *normative space* that confers an identity on certain entities and activities that could not exist – it would be logically impossible – if it were not situated within that space. Note that constitutive rules bring into existence new *normative* entities and activities, not new *physical* entities and activities. Or rather, constitutive rules do not bring into existence new *particular* things; rather, constitutive rules create particular functional *types*. These functional types – the roles, offices, positions, etc., that exist within rule-governed practices – can be, and usually are, *occupied by* or *instantiated by* physical things. This may allow us to say, of a particular thing, that in addition to being a physical object it is also a normative entity; but this is just a way of saying that it is a physical object with the power to function in certain ways within the normative space of a practice. This is a difference expressed in the way we describe, and therefore, interpret and understand particular (physical) things. For example, on the one hand, we give a physical description of an event by describing it as "a human

---

123 *[Note Added]* Searle here speaks of the collective *assignment* of a status. Though there may be circumstances in which this is an apt characterization of the source of a normative status – e.g. within the context of a new and emerging practice, or among the individuals responsible for the creation of a practice – I think it more accurate to say that status functions depend on collective *recognition* or *acceptance* by participants in the practice. Entering into most practices is a matter of training or initiation; it is a matter of, in some sense, submitting oneself to its constitutive rules. Think, for instance, of learning how to play chess: coming to understand the various permissible moves a type of piece makes certainly is not a matter of *assigning* these functions to the pieces.

124 Searle 2005, 7.

being running while carrying an object in an expanse of land from one spot to another." On the other hand, within the context of the game of football, we give a normative description of the same event by describing it as "a *receiver* scoring a *touchdown* on a *football field*." All of the items mentioned in this normative description presuppose the existence of the practice of the game of football. They are all physical things, which also have identities *as normative entities* by having a status within a practice.[125] Thus, to be a normative entity is to count as one; and to count as one is to be situated within the normative space created by a practice's constitutive rules.[126]

With this brief account of the nature of constitutive rules and practices on board, we can turn to collective agents. Collective agents are normative entities; they exist within the normative space of a practice. A football team, for example, is a collective agent that exists within the normative space of the practice that is the game of football. Similarly a parliament is a collective agent whose home is in the normative space of a particular (form of) government.[127] When playing their assigned part within the practice the individual agents that comprise collectives are also normative entities. This is not to say they are not real

[125] For a familiar example, notice this is just how it is with a language. Some physical things like sound waves and marks on paper are situated within the context of a language; thus, they are also meaningful *words* and *sentences*, or *nouns* and *verbs*, as well as *commands* and *greetings* and *insults*.

[126] One committed to an extremely austere physicalism may conclude from the fact that normative entities exist by virtue of having a status function in a practice that there are no such things. While there may be arguments in favor of such eliminativism, I am not going to address them here. In light of the fact of how few things would pass the ontological test on such a view, I take it to be a reasonable assumption that the burden of proof is on the person who wants to deny the existence of social and normative kinds.

[127] Notice that collective agents can also be considered practices themselves when compared to individual players and their teams. Like the team, the individual (types of) players – quarterback, running back, receiver, etc. – are also normative entities. They're situated within the practice of the game of football. We can also say they're situated within the more particular practice that is a team. Referees and linesmen are likewise situated within the practice of the game of football; they are not, however, parts of another practice within the game, namely, a team.

flesh-and-blood human beings: insofar as they act as players (quarterbacks, outfielders) or office holders (President, Chief Justice) they function in roles that arise from the constitutive norms of the practice. Like individual agents, collective agents have ends, goals the fulfillment of which is the aim of their actions. They pursue these ends, i.e. perform their actions, via the activities of the individuals who make them up.

Consider the following case: a quarterback throws a pass to one of the team's receivers; the receiver catches the ball and runs with it into the end zone; as a result of what these two individuals do, the team scores a touchdown and wins the game. What is significant about this example is the way it illustrates how a (formal) collective agent functions. Here we have a case of a collective agent (a football team) performing an action (scoring a goal, winning the game) by virtue of parts of the collective (individual players) doing certain things (throwing, catching, and running).

Now imagine a confused spectator who asks, "What is all the celebrating about?" "The team just scored the winning touchdown," you explain. He furrows his brow and responds: "What do you mean *the team* just scored a touchdown? I saw the same thing you did, and you know that is not what happened! The receiver scored the touchdown, not the whole team!" This person's confused, but he does have a point. If we understand the team to consist of all the players on the field, and if that is all we focus our attention on, then there is something to be said for the complaint. He is right that only the receiver brought the ball into the end zone. That is, if we are being very charitable, then we can imagine that

the novice spectator could think that in order for the team to score a touchdown every member of it has to pass into the end zone.[128]

What our bewildered spectator fails to understand is that *what it is* for a *team* to score a touchdown just is for a particular *player* playing a particular *position* to bring the ball into the end zone. Parts of collectives, in this case a member of a team playing a particular position, can act in ways that redound to the whole in which they are embedded because, like artifactual agents, collective agents are also teleologically constituted. What it means to say collectives are teleologically constituted is that the ends of a collective are determined by the constitutive rules of the practice, and the nature and function of the collective's parts depends on the ends the collective is designed to pursue. Thus the essential nature of a part of a collective – the functional role it plays in the whole – is determined by the essential nature of the collective itself – the ends it is designed to pursue as an agent. We can sum up this idea by saying that the collective's end is the *constitutive principle* of its parts; it is what enables us to understand their being constituted as they are, and therefore why they function as they do.[129]

The constitutive connection between the ends and aims of the whole collective and the nature of the roles or positions that are its parts grounds the power of some individual

[128] Obviously I am cheating a bit when I have this strange character failing to see how the team could score given only one player crossed into the end zone. I suspect that someone unable to perceive collective action like this would not grasp the concept of a team.

[129] Recall that in the previous section I said that the overall end or proper function of a robot – the jobs it is designed to perform – plays the same role in determining the nature and function of the robot's parts that the overall end(s) of a collective agent plays in determining the nature and functions of its part. The fact that in the robot case we speak quite literally about the parts of the agent, while in the collective case talk of parts is short hand for talking about the positions within the collective and their occupants, does not make a difference to the way in which the overall ends of these agents are their constitutive principles.

members of a collective *to function as the entire collective* on some occasions. For example, the constitutive rules of football determine that a football team is a collective agent that plays and trys to win football games; this is the team's collective end (at least while on the field),[130] which is determinative of its nature, the kind of thing it is. This top-down normative determination of essence and function is repeated at the level of particular players or positions; for what it is to occupy a certain position is to be authorized to perform certain activities that contribute to the team's collective end.

The essential connection between the overall aims of the collective and the constitution of its active parts is what enables the activities of some of these parts to redound to the entire collective. As we see in the case of the confused spectator, this is not because some parts of a collective become endowed with the physical power to, say, literally move all the members of the team into the end zone. Rather it is that within the normative space of a practice, the activity of the collective and its functional parts become more than just the physical movements that comprise this activity. These movements take on a normative/functional status, which transforms them from bodily movements of individuals to *moves* or *plays* of individuals occupying norm-governed *positions* within a practice. So it is not that some parts of a collective have a *physical power* to act for the collective; rather, it is

[130] I add this parenthetical remark because it is possible that the same individuals who make up a team can also comprise another collective agent off the football field. For example, perhaps the team that wins the game on Sunday afternoon consists of the same folks who put out fires as members of the volunteer fire department on Sunday evening. The football team and the fire department obviously have different ends as collective agents; thus, the selfsame individuals who make them up have different individuals ends when functioning in the normative space of the football team from the ones they have in the fire department. Brad might run the show as quarterback at the football stadium; but, as the junior most firefighter, he does yeoman's duty at the firehouse.

that parts of a collective have the *normative authority* to act on behalf of the collective. This is a matter of having the activity of the parts *count as* the activity of the whole within the normative space of the practice. Thus what accounts for the self-movement of a collective agent via its parts is the normative context in which this activity takes place. This is the context put in place by the constitutive norms of the practice, the norms responsible for the collective agent's existence.

3.7 Conclusion

The focus of this chapter has been on the standard story of self-movement, at the heart of which is the claim that self-movement consists in movements caused by a particular part or mechanism within an agent. The part or mechanism responsible for self-movement is the one whose proper function is to generate movements that are appropriately responsive to the agent's environment and that serve or constitute the agent's proper functioning. This relation between the proper function of the self-movement mechanism and the whole agent is one of teleological constitution, which is to say that the nature and function of the agent's parts depend on the overall ends and aims of the agent as a whole. This implies that the proper functions of the agent and its self-movement mechanism converge – the agent's proper function is to pursue certain ends and perform certain activities; the mechanism's proper function is to generate movements that contribute to the pursuit and achievement of these ends and constitute the performance of these activities. This intimacy between the agent's nature and proper function and the nature and function of the mechanism

responsible for self-movement allows us to say that movement caused by this mechanism is movement caused by the agent's nature. For that reason, movement caused by the mechanism redounds to the whole agent.

Thus, we can summarize my normative analysis of the standard story of self-movement as follows:

If A is a paradigmatic self-moving agent, then:

1. A has proper function F.
2. A is teleologically constituted: the nature and proper function of A's parts – A's constitution as a unified agent – depend on F.
3. When A's parts perform their proper functions, this contributes to the performance of F.
4. A's part P has the proper function of causing behavior that contributes to or constitutes F.
5. Given the essential connection between A's proper function (to perform F) and A's nature (A's constitution as a unified agent), P's proper function (to cause behavior contributory to or constitutive of F) is essentially connected to A's nature.
6. Behavior caused (in the right way) by P counts as movements caused by A's nature.
7. Behavior caused by A's nature count as movements caused by A (self-movement).

In Chapter 4 I extend this account to individual biological agents like us. In particular, the focus is on extending the concept of constitutive teleology to human beings in order to show how internally caused bodily-movements can count as movements performed by the whole agent. In order to make this case I will discuss an Aristotelian account of the nature of living things in which constitutive teleology plays an essential role. I will then identify the particular part in human agents whose nature allows it to function so that when it initiates and controls an agent's movements this is not just a case of an agent being passively moved by an internal cause; rather it counts as the agent purposively moving herself

FOUR

Desire and the Good

4.1 Introduction

In the discussion that follows I rely on insights from the previous chapters to articulate a response to PDA on behalf of CTA. I begin by retracing the path of the discussion up to this point.

In Chapter 1 I introduced the causal theory of action (CTA) as a philosophical account of what happens when someone acts. I argued that CTA is a naturalist theory of action because it tells a reductive story about human agency, one that reduces the process of a human being acting to the process of an agent's motivating beliefs and desires causing (in the right way) the bodily movements that comprise an agent's action. Some critics of CTA argue that it suffers from the problem of disappearing agents (PDA). PDA consists in the charge that CTA's reductionism effectively eliminates agents from their actions. This criticism is based on the idea that the bodily movements that constitute human actions are either brought about by agents themselves or caused by parts[131] of agents like mental states – not both. That is, proponents of PDA reject the idea that the phenomenon of whole agents

[131] As I have done throughout the discussion, I use 'part' here loosely to refer to elements of an agent's psychological or physiological makeup.

moving themselves can be reduced to the process of some parts of agents causing the movements of other parts.

In Chapter 2, I argued that philosophers have failed to distinguish two different ways of understanding PDA's concept of an agent playing an active role in her behavior. Following Frankfurt's influential work on the will, philosophers of action like Bratman and Velleman have interpreted PDA as a claim about CTA's ability to capture the involvement of an agent's higher rational faculties in its story of action. This interpretation of PDA takes for granted CTA's ability to capture basic self-movement that does not involve an especially robust exercise of rationality in its execution. As a result, it overlooks the fact that PDA also arises at the level of basic activity. The issue at this level is not whether CTA's belief-desire story of action can accommodate full-blooded human action – what Velleman referred to as human agency *par excellence* – but whether movements caused by beliefs and desires can be understood as movements performed by the agent herself. Thus, while sophisticated versions of CTA, like those defended by Bratman and Velleman, may successfully bolster CTA's ability to capture higher forms of rational agency, these accounts do not treat the more fundamental issue of how mental-state-caused bodily movements qualify as an exercise of an agent's basic capacity for self-movement. Since this issue about basic self-movement is prior to problems with CTA that arise at the level where actions involve exercises of practical rationality, the going attempts to deal with PDA are inadequate. This does not mean, however, that CTA has nothing to say in response to PDA.

The general shape of an answer to PDA emerges from Chapter 3's normative interpretation of the standard story of self-movement. Recall, PDA arises from the apparent incompatibility between the following two claims:

1. The movements that constitute intentional actions are caused by mental states.
2. The movements that constitute intentional actions are caused by agents.

These two claims seem incompatible because the agent herself is not identical to any of her mental states; nor can she be identified with any other mechanism within her with the power to move her limbs. The way to reconcile 1 and 2, and thereby save CTA from PDA, is to show how the movement of an agent by the agent can arise from movement of the agent by one or more of her parts. My normative interpretation of the standard story of self-movement allows for this.

At the heart of this interpretation is the idea that paradigmatic self-moving agents – either artifactual, conventional/collective, or organic – are teleologically constituted. What this means is that the nature and function of the agent's constitutive parts are essentially related to the agent's overall proper function, its characteristic ends, aims, and activities. If we accept the Aristotelian point that agents derive their nature from their function, that is, if we accept the idea that (at least part of) what it is to be a particular kind of active being depends on what things of that kind characteristically do, then we can understand teleological constitution as a state in which the function and the unified structure of an agent's parts can all be explained by the agent's nature, that which gives it an identity as the kind of thing it is. The beginning of a response to PDA, therefore, is that we can count

some behavior caused by certain parts of a teleologically constituted agent as self-movement because they are movements generated by a mechanism whose own nature is essentially connected to the nature of the agent as a whole. So although we cannot make sense of the notion of a whole agent being identical to one of her parts, the normative Aristotelian interpretation of the standard story forges an essential connection between the nature of the whole agent and the nature and function of the part that generates her behavior. Within the naturalistic-cum-reductive constraints of CTA, I suggest this is as close as we can get to capturing the notion of "agential origin" at the heart of our concept of human agency within a naturalistic event-causal framework.

So far the discussion has centered on artifactual and conventional agents, which is why at this point we have only the beginning of a response to PDA. The next step is to extend the account of the standard story of self-movement to organic agents like human beings. My purpose in this final chapter is to articulate an Aristotelian conception of the nature of living things, one that supports extending the concept of constitutive teleology to human agents. Where the concept of an agent's overall aims and proper functions plays a central role in understand the constitutive teleology – and, hence, the self-movement – of artifactual and conventional agents, the Aristotelian account of living agents centers on the concept of a creature's *good*.

4.2 An Aristotelian Conception of Living Things

In this section I layout the essentials of an Aristotelian[132] account of living things. According to the Aristotelian account, individual human agents, like their collective and artifactual counterparts, have teleological constitutions. They all embody a normative structure wherein the identity and function of their parts depend on the ends of the larger whole they comprise.

4.2.1 Aristotelian Forms

At the heart of Aristotle's discussion of living things is the concept of *soul*. It is important to point out at the outset that the Aristotelian understanding of the concept of the soul is entirely unlike the way in which the concept is generally understood in modern and contemporary philosophy. Our current understanding of the soul comes to us from Descartes, who used the term to refer to the locus of conscious mental life, the *res cogitans*. Thus the Cartesian concept of soul effects a division within a person between her animal, bodily nature, and her nature as a rational being. The term applies exclusively to the latter aspects of one's existence, thereby limiting the application of the concept to human and divine beings. The Aristotelian conception of the soul, on the other hand, is something that extends throughout the animal community, extending in fact to all living things. In fact, the soul is a particular instance of the more general category of *form*, a concept whose application

[132] It is important to keep in mind that the account of living things I develop below is Aristotel*ian*. That is, I am not claiming that the view necessarily matches the details of Aristotle's own view. What is important for my purposes is whether the view is a plausible and compelling account of living things; not whether the details of the view match precisely what Aristotle says. In other words, this is not a work of historical scholarship.

extends beyond the category of the living to substances of all kinds. So before turning to the souls of animals, I will begin with a brief discussion of Aristotelian forms.

According to Aristotle, substances or objects are composed of both matter and form. Something's matter consists of the material or parts out of which it is constituted. The form of a thing, on the other hand, is its structural and functional organization. As Korsgaard explains, an Aristotelian form is "the organization or arrangement of [a thing's] parts that allows it to be what it is, to do what it does, to do its job."[133] Implicit in this statement of form is the idea that, in addition to something's structure or architecture, form also involves the set of powers or capacities that are enabled by the nature of its material constitution. These criteria for form are most evident in the case of artifacts. Korsgaard uses the example of a house:

> For example the purpose of a house is to be a shelter, so the form of a house is the way the arrangement of the parts – the walls and roof – enables it to serve as a shelter. 'Join the walls at the corner, put the roof on top, and that is how we keep the weather out.' That is the form of a house.[134]

To be able to speak in terms of form we must be in a position to identify a particular function that is characteristic of a given object. In addition, the object itself must be composed of a complex arrangement of functioning parts that enables it to perform this function. (This means, for instance, that the concept of form does not apply to a rock or a pile of trash.) We satisfy these conditions for the application of the concept of form to a given object by situating it within the context of the kind to which it belongs. In fact, we

---

[133] Korsgaard 1996, 149.
[134] Korsgaard 1996, 107.

can say that something's form is what accounts for the fact that an object falls under a particular sortal like 'house'; and this just is having the material structure and functional capacities of a typical member of the kind. So talk of form and talk of kind- or sortal-dependence go hand-in-hand.

4.2.2 Animal Form – Species & Life-Cycles

This point about the kind-dependency of form facilitates the extension of Aristotle's concept of form to living things, for organisms are paradigmatic examples of things that belong to kinds, i.e. species. The Aristotelian claim is that the species membership of a given organism is what determines the organism's function and the nature and function of its parts. As Michael Thompson – a contemporary neo-Aristotelian – has emphasized, our judgments and descriptions of individual organisms make essential reference to the nature of the species or life form to which the creature belongs. A species, that is, is a "wider context" in which we situate a living creature, "a framework for interpreting the goings-on in … individual organisms."[135] In order to understand how this function talk is applicable to living things,

[135] Thompson, in fact, suggests that what it is to be alive is to be the proper subject of the type of judgment – "natural-historical judgment" – that is on hand when we interpret an organism and its activities in light of its species membership. (It is not necessary for my purposes here to take a stand on this point about what suffices for being alive.) What is special about this form of judgment, according to Thompson, is evident in the peculiar logical properties of the sentences we use to express them – what he refers to as "Aristotelian categoricals." For example, borrowing from Anscombe, we can make the following judgment about humans: "Human beings have 32 teeth." What is distinctive about this judgment, Thompson argues, is that it is tenseless and unquantifiable. In other words, we lose the meaning of the statement if we try to put it in the past or future tense, for it is not a statement about how human beings were or how they're going to be. Rather it is a claim, made in "a special kind of present tense", about the nature of the species *homo sapiens* (Thompson 2004, 49). Not only are such judgments somehow temporally restricted, they also have a kind of logical restriction; for we cannot go from the claim that humans have 32 teeth to the universal generalization that all

we need to note that corresponding to any given species is a typical form of life or life-cycle that characterizes the life of its members. According to Philippa Foot – another prominent neo-Aristotelian – when we speak of the life-cycle of a given species our subject is "how a kind of plant or animal, considered at a particular time and in its natural habitat, develops, sustains itself, defends itself, and reproduces."[136] In the case of most animals, at least, a part of this process involves perceiving, exploring, and successfully navigating its environment so as to procure nourishment, elude predators, and secure mates and appropriate shelter.[137] The concept of life-cycle, and the corresponding set of activities it comprises, is important for applying form to living things because it secures a place for talk of functions in the biological context.

Notice that being familiar with a life-form and its characteristic life-cycle brings with it the ability to make evaluative judgments about whether a given member of a species is healthy and well-functioning or defective and struggling to maintain itself as a living member of the species. So, for example, knowing that a prime source of nourishment for a cheetah comes from eating gazelle – a remarkably fast land animal – supports the judgment that a

---

humans have 32 teeth. The later claim is obviously false. It is not uncommon either to lose teeth or to be born without a full set of 32. Again, this is due to the fact that this Aristotelian categorical about the number of teeth human beings have is not a judgment of a particular member of the species; it is a judgment about the species itself. These two features of Aristotelian categoricals point to the fact that they are not descriptive statements; rather, they are expressions of norms or standards represented by the species and embodied in its typical or ideal or fully functioning members. See Thompson 1995.

[136] Foot 2001, 29.

[137] This is obviously an over-simplified picture of the typical life of many creatures, especially mammals and other high animals whose life-cycles involve activities like forming bounds and maintaining relations with other members of the species; learning and teaching behaviors involved in, say, grooming, hunting, and migrating; and, in the case of human beings at least, becoming appropriately initiated into and taking part in one's cultural milieu. The simpler life-cycle on which I am focusing in the text, however, is sufficient to convey the point I am making about the nature of living things.

cheetah that is unable to run at the speed typical of a well-functioning cheetah is deficient. For the cheetah's inability to do something that is typical of its kind – namely, run very fast – precludes it from procuring sustenance. An invalid cheetah, therefore, is not likely to last very long. Now there are likely a number of reasons why a given cheetah is unable to attain the kind of speed necessary to pursue its quarry, but they will all involve a deficiency in some element within the animal's physiology, that is, within the organized structure of its bodily organs and systems.[138] This point allows us to see the connection between an animal's function and the functional organization of its parts that is at the heart of the Aristotelian concept of form. It also highlights the essential connection between the nature or identity of an animal and its proper functioning; for, as in my cheetah example, the inability to partake in the activities that are essential to the life-cycle of an animal's kind eventually results in the animal's destruction.

In the discussion so far, I have attempted to trace a path from the Aristotelian concept of form to the biological notions of species and life-cycle. Before moving to an explicit statement of Aristotle's concept of animal soul, I want to draw out a number of ideas or connections among concepts that this effort has uncovered. The first is a general connection between form and kind-membership; to be *enformed* involves being a certain type of thing, belonging to a particular sort or class. The second connection is between being a member of a particular class and having a particular function and material structure. As we saw, to be a *house*, to belong to that class or category, is to be a unified collection of parts

[138] This, of course, does not preclude the ultimate cause of the defect being an environmental factor. Such factors cannot deprive an animal of its ability to properly function without affecting its physiology in some way.

constructed in such a way that it can adequately provide shelter. These first two connections fall out of the general picture of Aristotelian form. Their application to living things involves translating talk of kinds and functions into talk of species and life-cycles. Thus, the third connection is between being a member of a given species and having a particular life-cycle, a certain active process of development and self-maintenance that characterizes the typical life of a bearer of the life-form. I claimed that we can understand the function of an organism as the successful navigation through the life-cycle endemic to its kind. I also noted that the concepts of species and the typical life-cycle of a kind of organism are connected to the normative evaluation of specific creatures, to judgments about the degree to which they are flourishing or floundering. So being a healthy, flourishing, well-functioning participant in the life-cycle of a species is performing the proper function of the species. Finally, in order for an animal to maintain its nature, to retain its identity as a member of its species, its ability to participate in its life-cycle must be on-going. The life of an animal, therefore, has an inherent circularity, for it consists in performing those activities constitutive of its life-cycle to the right degree of success such that the animal can continue to perform them.[139]

We can conclude from the above reflections that there is an inherent connection between form (species), function (life-cycle), goodness (proper functioning), and nature (identity). Being an animal essentially involves belonging to a certain species. Therefore, an animal's nature and identity depends on its active participation in the activities that comprise the life-cycle endemic to its kind. It is only by functioning in this way that an animal can

[139] Korsgaard (1996, 49) captures this idea when she describes living things as having "self-maintaining" forms.

maintain the flourishing state of health or full-functioning that supports its continued existence. Thus, an animal performs its function, and thereby maintains itself as the kind of thing it is, by pursuing and (to the necessary degree) achieving its *good*. This connection between an animal's good, its proper function, and its nature or identity brings us to the doorstep of the Aristotelian concept of soul.

4.2.3 Aristotelian Souls & Constitutive Teleology

We can now turn to the Aristotelian view of soul, i.e. the form of living things.

Just as the form of a house is the organization of its parts that enables it to serve as an adequate shelter – to perform the proper function of a house – so too the soul of a living thing is the species-dependent functional arrangement of its parts that allows it to successfully participate in the life-cycle of its species. As I noted earlier, there is an implicit reference in this appeal to the functional arrangement of an animal's parts to the set of powers and capacities that are enabled by having this type of material constitution. In fact, many articulations of Aristotle's concept of soul emphasize these active abilities. For example, James Wallace writes, "the *psuche* [i.e. soul] of an organism is conceived as the capacities and tendencies it possesses to carry on the activities characteristic of the mode of life of the creature's kind."[140] J. L. Ackrill, on the other hand, provides an account of soul that focuses on both structure and function:

[140] Wallace 1978, 25-26.

> To say that this body is *alive* is, according to Aristotle, to say that it has powers of a certain kind … To speak of soul, then, is to speak of the ability to do certain things … [H]aving a certain kind of soul … makes something a plant, or an animal, or a man. Having a soul is precisely what makes this collection of flesh, bones, etc. an animal – and *one* animal – just as the shape and structure are what makes some timber *one* thing, namely a table.[141]

Both Wallace and Ackrill make clear a point I stressed earlier, namely, that we ought not interpret Aristotle's concept of soul in traditionally Cartesian terms. This is because "a soul is not an 'it' housed in the body, but a functional structure in and of matter."[142] Thus, allowing a role for Aristotle's notion of soul does not entail any dualist or otherwise non-naturalist commitments. An Aristotelian can accept that all living things are wholly composed of physical stuff, and that, therefore, for any type of thing there is a complete underlying story about what goes on at the physical level. What the view is committed to is a rejection of a form of reductionism that says that the properties and capacities of things are reductively identifiable with and completely explicable in terms of the laws governing their physical constituents, those of physics and chemistry. An Aristotelian can accept that countenancing non-material entities or properties, those whose existence would require the violation of the laws of physics, is something completely foreclosed by respect for the natural sciences without eschewing the causal-explanatory power of appeals to higher-level phenomena. As Michael Frede explains, "to say that an object has a certain nature [i.e. form] is not to postulate a mysterious force or a mysterious kind of causation; it is to say something about how the object and its behavior have to be understood and to be

[141] Ackrill 1981, 70.
[142] Nussbaum and Putnam 1992, 56.

explained." The appeal to form, that is, "adds a further level of understanding to what happens." [143]

Recall that my purpose in articulating this Aristotelian conception of the nature of living things is to provide a basis upon which I can extend the normative account of self-movement I developed in the last chapter to organic agents like human beings. At the heart of that account is the concept of constitutive teleology, the idea that the nature and function of the parts of (paradigmatic) self-moving agents depends on the overall aims, ends, or functions of the agent. What should now be clear is that, according to the Aristotelian view, living things are also teleologically constituted. The Aristotelian view maintains that the organic parts comprising an animal are functionally arranged so that the animal can exercise the powers and capacities needed to successfully pursue its natural ends or perform its natural functions. Now so far, I have focused on the idea of the "functional arrangement" of something's parts in my discussion of the Aristotelian view, but it is important to see that this notion covers more than how the parts fit together. This points to a way in which Korsgaard's house example is may be misleading when it comes to understanding the nature of animals, for she explicitly articulated the point in terms of how the walls and the roof are *joined* and *put together*. This makes it easy to overlook that *the nature of the walls and the roof themselves* matter to the ability of the structure to serve the function of a house. Returning to living things, talk of the functional arrangement of an animal's parts also refers to the particular parts themselves. Just as the way in which an organic part fits into the overall

[143] Frede 1992, 102.

structure of an animal is a fact about the animal's soul, so too is the part's own nature and function related to the form of the animal. As we have seen, this translates into the claim that the nature and function of the organic parts of an animal depend on the role these parts play in the animal's performance of its proper function, i.e., in the active pursuit of its good. Therefore, like the functions a robot is designed to perform, and like the ends a collective agent is created to achieve, an animal's *good* is its *constitutive principle*.

Before moving on, it is important to note that the question of the source of constitutive teleology for organic agents is an enormously difficult one. Some answers appeal to the process of evolution while others point to the intentions of an intelligent designer. It is not my purpose here to take a stand on the debates surrounding this issue, but rather to show what work can be done if we take on the Aristotelian claim that, as with other substances or objects, organic agents are teleologically constituted.

4.3 An Evaluative Conception of Desire

According to the normative version of the standard story of self-movement I articulated in Chapter 3, behavior generated by an internal cause can count as the self-movement of the whole agent when the mechanism responsible for the behavior is essentially connected to the agent's nature. The Aristotelian conception of living things I have been discussing provides the groundwork for the application of this account to human agents. The next step is to identify the particular mechanism in human agents that qualifies as the proper internal cause of self-movement. Given a certain traditional picture of desire, we can understand

desires and the faculty of desire more generally as this proper internal cause. Thus CTA's reliance on desires in its story of what happens when someone acts, when combined with the Aristotelian view, yields a response to PDA.

According to what I will call an *evaluative conception of desire*, desires are causally efficacious states whose proper function is to track an agent's good and put her in pursuit of it. This differs from other accounts of the nature of desire that connect desire and the good. The most prominent alternative version of an evaluative account of desire maintains that all desires are for the good, or for the appearance of the good. This is the view expressed by the claim that the ends of human action are desired *sub species boni*, i.e. under the guise of the good. On this view, to desire something necessarily involves seeing it as in some sense good or as in some way conducive to one's well-being. The problem with this view is that it falters in the face of our experience as agents. That is, the life of a typical agent involves many instances of irrationality in which one desires something that one knows is not good. This happens every time one reaches for that second (or third…) piece of cake, or when one pours that supposed last drink. We are all too familiar with wanting what we know we should not. The "guise of the good" view of desire gets the phenomenology of moral agency wrong; that is, it misidentifies the true nature of much of our experience as agents. This is Velleman's point in the following:

> A tendency to desire things under negative descriptions is an essential element of various emotions and moods such as silliness, self-destructiveness, or despair. A mood of playfulness is, in part, a disposition to form desires for things conceived as having no particular value; a self-destructive mood is, in part, a disposition to form desires for things conceived as harms; and so on. None of these [very common]

> desires could retain its characteristic idleness or perversity if it involved an attempt at getting things [about the good] right.[144]

So it is important to keep in mind that the evaluative conception of desire that I recommend does not claim that to desire something is necessarily to see it as good or to pursue it under the guise of the good. My view is that properly functioning desires are states that track the agent's good and put her in pursuit of it. This accommodates the common cases of desiring the bad because it allows for the possibility of malfunctioning desires.

4.4 The Faculty of Desire

This brings us to the crux of my argument. CTA can respond to PDA by combining the Aristotelian conception of the nature of living things with this evaluative conception of desire. According to the Aristotelian view, living agents are teleologically constituted which entails that the function of the organic parts is to promote the overall good of the organism they comprise. According to the evaluative conception of desire, desires are behavior-generating states whose proper function is to track the good and put her in pursuit of it. Thus the natures of both the agent herself and her faculty of desire are essentially related to the human good. This is why desire-caused bodily movements (caused in the right way) count as behavior the agent – understood as the whole living organism – causes instead of being movements only ascribable to one of the agent's parts. These behaviors are equivalent

[144] Velleman 2000, 118.

to self-movement because they originate from an element of this active self whose essential nature and proper function is essentially connected to the agent's own nature.

I need to clarify my use of the term 'desire' by drawing a distinction between two senses of the word. We can use the term 'desire(s)' to refer to the *behavior-generating states* that play a role in CTA's story of action; or we can use 'desire' to refer to what Aristotle calls *the faculty of desire*.[145] CTA uses 'desire' in the first sense in its account of intentional action – desires as motivational *states* that proximately cause self-movement. In order to provide an account of why these states are properly considered the causes of self-movement, however, we need to bring in the broader idea of a *faculty* of desire in which these states originate and from which they exercise their causal powers.[146]

We can understand the faculty of desire as analogous to the motor-control mechanism (MC) found in the kind of robots discussed in Chapter 3. Like the robot's MC, an animal's faculty of desire is responsible for generating behavior that is appropriately responsive to the creature's environment. Thus, just as the MC is sensitive to input from its various environmental sensors, so the faculty of desire is responsive to input from the environment via sense perception. The robotic MC and the faculty of desire respond in similar ways to the data they receive. While the MC sends particular commands or signals to

[145] Where I speak here of "faculty" of desire, one can also speak in terms of a desiderative "power" or "capacity".

[146] For the purposes of this discussion we do not need a fully worked-out understanding of the faculty of desire, for that would require taking a stand on issues beyond this discussion. Following Aristotle we can say that the faculty of desire goes hand-in-hand with having the capacity to perceive one's environment; and, thus, the faculty of desire involves the power to initiate movement in response to perception. For our purposes, this will suffice as a basic idea of what I mean when I speak of the faculty of desire. As I go on to explain, the faculty of desire, as I understand it, also involves the ability to form particular desires for specific practicable goods (e.g. a desire for *this drink*) that derive from more general desires.

motors that move its arms, wheels, legs, etc., the faculty of desire generates particular motivational states (desires) responsible for causing behavior. Finally, the concept of proper function applies to both the MC and the faculty of desire. The MC functions properly when it enables the robot to effectively navigate its environment, thereby properly pursuing the ends and aims the robot as a whole is designed to achieve. The proper function of the faculty of desire is to regulate the formation of desires and the exercise of their power to cause bodily movements so the resulting behavior is aimed at the agent's good, i.e., behavior that contributes to or constitutes successful participation in the life cycle of its species. This implies that the proper function of a particular motivating desire that arises out of this faculty is to put the agent in pursuit of particular goods achievable in a given context of action that have the right contributory relation to her overall good.

I suggest we understand how the faculty of desire could perform this function in the following way. At the heart of the faculty of desire is a set of fundamental desires to pursue certain kinds of ends and to participate in certain forms of activity that are instrumental to or constitutive of the animal's good (successful participation in its life cycle). These basic or fundamental desires drive the faculty of desire's operations by instilling within the agent a natural inclination toward what is good for it and by motivating and structuring the processes through which particular desires are formed and transmit their causal powers. So, while the basic desires are not for the agent's good *qua* good, they are desires for what in fact constitutes the agent's good.

To illustrate, consider a primitive animal agent – call her Prima – whose life form has a very simple life cycle, one that involves performing basic life functions like eating and drinking, resting, having sex, and eluding danger. Prima achieves her good when, through her active engagement with her environment, she partakes in these activities. Given the nature of her good, and given the fact that the proper function of her faculty of desire is to put her in pursuit of it, at the heart of Prima's faculty of desire are basic desires for sustenance, security, and reproduction. By instilling in her a basic tendency to be attracted to and to pursue ends achievement of which constitutes a flourishing condition, these basic desires effect a natural orientation of Prima's faculty of desire in the direction of her good. Prima would be severely defective if she lacked these desires because, without such tendencies towards her good, her faculty of desire cannot function properly; it cannot regulate the processes through which her particular here-and-now desires form and lead to action in a way that is appropriately sensitive to what she needs to flourish.

To recap: the faculty of desire regulates the processes through which desires form and exercise their power to cause behavior. This regulative function is driven by a set of basic desires for certain ends and activities constitutive of successful participation in the life cycle of the agent's species. Of course, an agent's good just is successful participation in the life cycle of its species. Thus, the faculty of desire is functioning properly when its functioning results in the agent desiring the pursuit of a good life by desiring particular ends and activities constitutive of that life. The particular desires for these ends and activities themselves function properly when they cause (in the right way) the appropriate bodily

movements (behavior) whereby these good ends are achieved and good courses of action are pursued.

This point about the intentional content of the desires at the heart of the faculty of desire is important; for if we could not make sense of the faculty of desire's intrinsic orientation to the good without positing a foundational desire whose content is something like "to pursue the good", then it would be difficult to take seriously the idea that this kind of faculty can be found in non-rational animals. Though there is much debate about the mental lives of non-linguistic or non-concept-using animals, most accept that something like the concept of desire is applicable to a broader swath of the animal kingdom. At the same time, most agree that a concept like "goodness" likely goes beyond the mental capacities of non-human animals. It would be fatal for my view is what I say about the faculty of desire could not be extended to non-human animals because I am arguing that CTA is an instance of the standard story of self-movement, a story that obviously accounts for more than human movement – something made evident by my discussions of robots. [147]

[147] This is not to say that a desire explicitly for the good *qua* good cannot play a role in the functioning of the faculty of desire, by, for example, motivating other behavior-generating desires. My claim is that such a desire need not be hard-wired, as it were, into the faculty of desire in order for the faculty to track the good. The reason this matters is that I do not want to give an account of the faculty of desire that restricts it simply to human or rational agents, those agents we can comfortably understand as possibly having a desire whose own content is that agent's good. In other words, while I think we can accept that the basic desires I discuss are present in some form throughout the animal kingdom (though not necessarily present in every type of animal – I am agnostic about, say, earthworms), I do not think the same can be said for the desire for one's good.

4.5 The Digestion Problem Revisited

With this account of how the faculty of desire works and of the way in which the faculty is connected to an agent's good, I want to return to my discussion in Chapter 3 (section 3.4) of the digestion problem. Recall the problem with (A_1) was that it was over inclusive; it counted some movements that agents do not actively performed as instances of self-movement. This was due to the fact that (A_1)'s "proper function criterion", as I called it, stated that movements that contribute to the proper function of the agent as a whole count as self-movement. This had the awkward implication that movements of an agent's digestive track, or the numerous internal goings-on inside a mobile robot, count as self-movement in the same way that movements involved in walking or eating do. I appealed to an ethological conception of behavior to overcome this worry. This resulted in a new criterion for the proper cause of self-movement:

(A_1)* The proper cause of self-movement is the mechanism whose proper function is to generate behavior – movements that occur when agents perform the activities that characterize their normal interactions with their environment – that constitutes or contributes to the proper function(s) of the agent as a whole.

When I articulated (A_1)*, I made clear that the account of behavior on which it rests was not intended as a complete analysis of the concept. This allows, therefore, for the possibility of movements counting as behavior that do not seem to be the kinds of movements that we would ascribe to the agent as a whole. I want to end this dissertation by

returning to the digestion problem because I think my Aristotelian account of self-movement has the resources to say that, a least in the case of self-moving agents like us with a faculty of desire, this possibility is foreclosed.

There is a direct connection between desire[148] (both faculty and state) and the human good that differentiates it from other organic parts and processes. Desire is intrinsically and directly related to the agent's good by virtue of the fundamental desires that drive of the faculty of desire. Because these basic desires are intentional states, an essential feature of them is that they are directed to something beyond themselves. As I explained above, the desires driving the faculty of desire are for the basic elements (ends and activities) constitutive of the agent's good. Thus, the desires driving the faculty of desire are essentially directed to the agent's good. We cannot give a full specification of the nature of these states without explicitly invoking the good of the agent of whom they are states. This is evident in the following table:

Organic System		*Function*
Digestion	*contributes to the animal's good by*…	extracting nutrients from food.
Excretion	*contributes to the animal's good by*…	expelling waste.
Respiration	*contributes to the animal's good by*…	oxygenating blood.

[148] Unless otherwise noted, I will use the singular term 'desire' to refer to the faculty and its states as a whole.

Now consider the specification of the function of desire:

Desire	*contributes to the animal's <u>good</u> by…*	causing behavior in pursuit of its <u>good</u>.

Using this table, we can further highlight the distinctive and direct connection between desire and the human good – as opposed to the indirect connection between the good and the other elements of an agent's teleological constitution – by comparing what happens when desire malfunctions with cases in which these other organic systems fail to work properly.[149] Say, for instance, I really want that sixth and final slice of pizza, so I grab it and gobble it down. I had major indigestion after the third slice, so this sixth one is bound to do me in.

Let us compare how my faculty of desire and my digestive system work in this case. Given that eating six slices of pizza is unhealthy no matter how you slice it (so to speak), and given that I am already suffering from indigestion, it is clearly not good for me to eat the sixth slice. Thus, when I crave the slice and reach out and grab it my faculty of desire is not functioning properly. It is not properly tracking my good, and it causes me to move in such a way that I go after something that is obviously bad for me. This last point – that it is obviously bad for me – is important, because it points to the fact that the malfunction is not somewhere else, like in my olfactory mechanism or in my stomach. For the smell of the pizza makes me slightly nauseated, and my stomach feels like it is about to explode. So my

[149] I owe the following example to Mark Lance.

faculty of desire is clearly malfunctioning, and this is bad for me. That is, when it comes to being a healthy, well-functioning human being I am very much off the mark. I am simply not pursuing my good, and thus I am failing to perform my function *tout court*. Turning to my digestive system, we can now ask how its functioning in this situation affects my good. First let us assume that my digestive system is properly functioning, and as a result I successfully digest the sixth slice of pizza. This means that I am successfully absorbing large quantities of grease and fat that will cause me to collapse in pain. What this shows is that my digestive system can perform its proper function without the results of this performance being part of my good. In fact, this is a case in which it would have been better for me had my digestive system malfunctioned. But unlike in the case of my faculty of desire, if my digestive system failed to digest the food it would be malfunctioning, but I, qua whole human animal, would not be. Malfunctions involving my digestive system, therefore, do not result in my going completely astray in the pursuit of my good. That is, they do not result in my failing to function properly *tout court* as a human animal.

The lesson to draw from this comparison between the faculty of desire and the digestive system is the following: the proper function of the faculty of desire involves being *differentially responsive* to the good. The faculty of desire has the power to be differentially responsive in this way by virtue of the fact that, via the representational content of desires, it can take in information about what is at stake for the agent's good in a given context of action. In other words, the faculty of desire is differentially responsive to the good because it can track the good; and it can track the good because it can represent the good. Other

organic mechanisms or faculties or processes lack the power to differentially respond to the good, therefore, because they cannot intentionally represent it. What follows from this is that even if we count movements caused by other organic parts as *behavior* it still does not qualify as self-movement. Self-movement, movements actively performed by the entire animal, consists of behavior caused by the part of an agent whose proper function is to cause behavior differentially oriented to the agent's good. I have argued that this mechanism is the faculty of desire.

BIBILIOGRAPHY

Ackrill, J. L. 1981. *Aristotle the Philosopher*. Oxford: Oxford University Press.

Alvarez, M. and Hyman, J. 1998. "Agents and their Actions." *Philosophy* 73: 219-245.

Anscombe, G.E.M. 2001. *Intention*. Cambridge, Mass.: Harvard University Press.

Aristotle. 1984. *The Collected Works of Aristotle*. Barnes, J. (ed.). Princeton: Princeton University Press.

Bayne, T. and Levy, N. 2006. "The Feeling of Doing: Deconstructing the Phenomenology of Agency." In Sebanz, N. and Prinz, J. (eds.). *Disorders of Volition*. Cambridge, Mass.: MIT Press.

Bishop, J. 1989. *Natural Agency*. Oxford: Oxford University Press.

Bittner, R. 2001. *Doing Things for Reasons*. Oxford: Oxford University Press.

Boler, J. F. 1968. "Agency." *Philosophy and Phenomenological Research* 29: 165-181.

Brandom, R. 1997. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, Mass.: Harvard University Press.

Bratman, M. 2001. "Two Problems About Human Agency." *Proceedings of the Aristotelian Society* 101: 309-326.

Brooks, R. 1991. "Intelligence Without Representation." *Artificial Intelligence Journal* 47: 139-159.

Burge, T. 1993. "Mind-Body Causation and Explanatory Practice." In Heil and Mele 1993.

Child, W. 1994. *Causality, Interpretation, and the Mind*. Oxford: Oxford University Press.

Collins, A. W. 1986. *The Nature of Mental Things*. Notre Dame, Ind.: Notre Dame University Press.

Coope, U. 2007. "Aristotle on Action." *Proceedings of the Aristotelian Society Supplementary Volume* 81: 1-30.

Dancy, J. 2001. *Practical Reality*. Oxford: Oxford University Press.

Davidson, D. 1993. "Thinking Causes." In Heil and Mele 1993.
____ 1980. *Essays on Actions and Events*. Oxford: Oxford University Press.

Davis, W. 2005. "Reasons and Psychological Causes." *Philosophical Studies*, 122, pp. 51-101.

Dretske, F. 1988. *Explaining Behavior*. Cambridge, Mass.: MIT Press.

Dreyfus, H. 1992. *What Computers Still Cannot Do: A Critique of Artificial Reason*. Cambridge, Mass.: MIT Press.

Enç, B. 2005. *How We Act*. Oxford: Oxford University Press.

Foot, P. 2001. *Natural Goodness*. Oxford: Oxford University Press.

Foran, S. 1997. "Animal Movement." Doctoral Dissertation, UCLA.

Frankfurt, H. 2002. "Reply to Michael E. Bratman." In Buss, S. and Overton, L. (eds.). *The Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, Mass.: MIT Press.

____ 1988. "The Problem of Action." *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frede, M. 1992. "On Aristotle's Conception of the Soul." In Nussbaum and Rorty 1992.

Freeland, C. A. 1994. "Aristotle on Perception, Appetition, and Self-Motion." In Gill, M. L. and Lennox, J. G. (eds.). *Self Motion: From Aristotle to Newton*. Princeton: Princeton University Press.

Furley, D. J. 1980. "Self-Movers." In Rorty, A. O. (ed.). *Essays on Aristotle's Ethics*. Berkeley, Calif.: University of California Press.

Gustafson, D. 1981. "Passivity and Activity in Intentional Action." *Mind* 90: 41-60.

Haldane, J. 1998. "A Return to Form in the Philosophy of Mind." *Ratio* 11: 253-277.

Haugeland, J. 1998. *Having Thought: Essays in the Metaphysics of Mind*. Cambridge, Mass.: Harvard University Press.

Hauser, L. 1994. "Acting, Intending, and Artificial Intelligence." *Behavior and Philosophy* 22: 22-28.

Heil, J. and Mele, A. (eds.). 1993. *Mental Causation*. Oxford: Oxford University Press.

Hobart, R. E. 1934. "Free Will as Involving Determinism and Inconceivable without It." *Mind* 43: 1-27.

Horgan, T. 2007. "Mental Causation and the Agent-Exclusion Problem." *Erkenntnis* 67: 183-200.

Hornsby, J. 2005. "Agents and Actions." In Hyman, J. and Steward, H. (eds.). 2005. *Agency and Action*. Cambridge: Cambridge University Press, 2005.

_____ 2004. "Alientated Agency." In De Caro, M. and MacArthur, D. (eds.). 2004. *Naturalism in Question*. Cambridge, Mass.: Harvard University Press.

_____1997. *Simple-Mindedness: In Defense of Naïve Naturalism in the Philosophy of Mind*. Cambridge, Mass.: Harvard University Press.

Irwin, T. H. 1990. *Aristotle's First Principles*. Oxford: Clarendon Press.

Jacobs, J. 1984. "Teleology and Essence: An Account of the Nature of Living Things." *Nature and System* 6: 15-32.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

_____ 1998. *Mind in a Physical World*. Cambridge, Mass.: MIT Press.

Korsgaard, C. 2002. "Autonomy, Efficacy, and Agency." In *Self-Constitution: Action, Identity, and Integrity*. http://www.people.fas.harvard.edu/~korsgaar/Korsgaard.LL3.pdf.

_____ 1996. *The Sources of Normativity*. Cambridge: Cambridge University Press.

Kraut, R. 2007. *What Is Good and Why: The Ethics of Well-Being*. Cambridge, Mass.: Harvard University Press.

Lear, J. 1988. *Aristotle: The Desire to Understand*. Cambridge: Cambridge University Press.

Lycan, W. 1981. "Form, Function, and Feel." *Journal of Philosophy* 78: 24-50.

McDowell, J. 1998. "The Content of Perceptual Experience." In *Mind, Value, and Reality*. Cambridge, Mass.: Harvard University Press.

____ 1997. "Reductionism and the First Person." In Dancy, J. (ed.). *Reading Parfit*. Oxford: Blackwell.

____ 1994. *Mind and World*. Cambridge, Mass.: Harvard University Press.

Melden, A. I. 1961. *Free Action*. London: Routledge & Kegan Paul.

Mele, A. 2000. *Motivation and Agency*. Oxford: Oxford University Press.

Moravcsik, J. 1994. "Essences, Powers and Generic Propositions." In Scaltsas, T., Charles, D., and Gill, M. (eds.). *Unity, Identity, and Explanation in Aristotle's Metaphysics*. Oxford: Oxford University Press.

Murphy, M. 2001. *Natural Law and Practical Rationality*. Cambridge: Cambridge University Press.

Nagel, T. 1969. "The Boundaries of Inner Space." *Journal of Philosophy* 66: 452-458.

Nussbaum, M. and Putnam, H. 1992. "Changing Aristotle's Mind." In Nussbaum and Rorty 1992.

Nussbaum, M. and Rorty, A. O. (eds.). 1992. *Essays on Aristotle's De Anima*. Oxford: Oxford University Press.

Pavlopoulos, M. 2003. "Aristotle's Natural Teleology and the Metaphysics of Life." *Oxford Studies in Ancient Philosophy* 24: 133-181.

Rawls, J. 1999. "Two Concepts of Rules." In Freeman, S. (ed.). *Collected Papers*. Cambridge, MA: Harvard University Press.

Richardson, H. 1992. "Desire and the Good in *De Anima*." In Nussbaum and Rorty 1992.

Schueler, G. F. 2003. *Reasons and Purposes: Human Rationality and the Teleological Explanation of*

Action. Oxford: Clarendon Press.

Schroeter, F. 2004. "Endorsement and Autonomous Agency." *Philosophy and Phenomenological Research* 69: 633-659.

Searle, J. 2005. "What is an Institution?" *Journal of Institutional Economics* 1: 1-22.

_____1995. *The Construction of Social Reality*. New York: Free Press.

Sehon, S. 2005. *Teleological Realism*. Cambridge, Mass.: MIT Press.

Steward, H. 1997. *The Ontology of Mind: Events, Processes, and States*. Oxford: Clarendon Press.

Strawson, P.F. 1985. "Causation and Causal Explanation." In Vermazen, B and Hintikka, J. (eds.). *Essays on Davidson: Actions and Events*. Oxford: Oxford University Press.

Taylor, R. 1966. *Action and Purpose*. Englwood Cliffs, N.J.: Prentice Hall.

Thompson, M. 2004. "Apprehending Human Form." In O'Hear, A. (ed.). *Modern Moral Philosophy*. Cambridge: Cambridge University Press.

____ 1995. "The Representation of Life." In Hursthouse, R., Lawrence, G., Quinn, W. (eds.). *Virtues and Reasons*. Oxford: Clarendon Press, 1995.

Velleman, D. 2000. *The Possibility of Practical Reason*. Cambridge: Cambridge University Press.

Wallace, J. 1978. *Virtues and Vices*. Ithaca, N.Y.: Cornell University Press.

Wallace, R. J. 1999. "Three Conceptions of Rational Action." *Ethical Theory and Moral Practice* 2: 217-242.

Wilson, G. 1988. *The Intentionality of Human Action*. Palo Alto, Calif.: Stanford University Press.

Wilson, R. 2005. "Persons, Social Agency, and Constitution." *Social Philosophy and Policy* 22: 49-69.

____ 1995. *Cartesian Psychology and Physical Minds: Individualism and the Science of Mind.* Cambridge: Cambridge University Press.

Yaffe, G. 2000. *Liberty Worth the Name: Locke on Free Agency*. Princeton: Princeton University Press.