UNIVERSITY OF CALIFORNIA
RIVERSIDE

Shared Agency in the Kingdom of Ends: Toward a Social Constitutivism

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Philosophy

by

Zachary Charles Bachman

June 2018

Dissertation Committee:
Dr. Peter Graham, Chairperson
Dr. Andrews Reath
Dr. Michael Nelson
Dr. Luca Ferrero

The Dissertation of Zachary Charles Bachman is approved:

_____

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

My graduate education has been a long and grueling process. Including my time at Texas A&M where I got my master's degree, I've been grinding it out in the minor leagues of academia for nine years. During that time, I've accrued a lot of debts, both professional and personal.

As my dedication indicates, I am greatly indebted to my parents. I've spent almost all of my 20s and half of my 30s on my post-secondary education. They never once asked me when I was going to get a real job. In fact, when I was applying to PhD programs while I was at Texas A&M I asked my mother, who is herself a lawyer, whether I should instead be applying to law schools. She told me that I absolutely shouldn't. All of her colleagues, she said, were jealous of what I was doing and that I should ride it out as long as I could. I don't know if that was the best advice, but it was very supportive, and I am grateful for that.

I have four sisters with whom I text almost every day. Graduate school can be lonely and soul-crushing at times. There have been days where I felt like I was surely never going to finish and that I had wasted a decade of my life on a foolish dream. My sisters always reminded me that they would love me no matter what happened with my academic career. They reminded me that the most important thing is life is family. Whatever happens with my career, I will always have them.

To properly acknowledge all of the people I am indebted to would really require writing acknowledgements of more than 20 pages. Rather than do that, I will do what

most people do, which is just list all of the people I am indebted to. I will try to find a proper way to thank them later.

This dissertation grew out of important conversations I had over the years with many colleagues in workshops, reading groups, during happy hour at the Getaway, and just hanging out in our offices on the third floor of the HMNSS building and the SAR room. (Yes, I know 'room' is redundant there) My greatest debts are to David Beglin and Meredith McFadden, both of whom are very close friends and have come to understand me, both philosophically and personally, better than perhaps I do myself. They have helped me become not just a better philosopher, but also a better person. (The three of us are now done! Huzzah!) I am also indebted to Justin Coates, Taylor Cyr, Jorgen Hansen, Andrew Law, Patrick Londen, Andrew MacDonald, Max McCoy, Benjamin Mitchell-Yellin, Max Murphy, Jonah Nagashima, Jeremy Pober, Patrick Ryan, Avery Snelson, Will Swanson, Yvonne Tam, and Monique Wonderly.

I am also, of course, indebted to my dissertation committee, Peter Graham, Andrews Reath, Michael Nelson, and Luca Ferrero. I have learned quite a lot from you all over the years. I just wish my dissertation did a better job of representing this fact.

I am also indebted to several other faculty members at UCR who weren't on my dissertation committee: Maudemarie Clark, Carl Cranor, John Martin Fischer, Pierre Keller, Coleen Macnamara, Jozef Müeller, Erich Reck, Eric Schwitzgebel, Howie Wettstein, and Mark Wrathall. You all supported me in various ways, sometimes with comments on written work, sometimes with arguments during reading groups, sometimes with words of encouragement, sometimes with tips on teaching, sometimes simply by

showing interest in my work, other times by finding ways to secure funding for me. UCR was really a great place to study philosophy. Thank you.

I am grateful to Lisa the human and Alfred the cat for making my time away from my work so enjoyable. You bring levity to my life. I look forward to the day when we will again be living under the same roof. At least now I will only have one more hour in the day than you.

I should also add that I am grateful to the philosophy program at Sam Houston State University for hiring me. If they didn't hire me I would be looking for adjunct work instead of finishing my dissertation.

Finally, I am grateful to Springer for permission to use previously published work. Chapter 2 of this dissertation contains elements of my essay "Moral Rationalism and the Normativity of Constitutive Principles" which originally appeared in *Philosophia* 46. I am also grateful to an anonymous reviewer for that journal who gave me excellent comments that lead to a much better paper, and who eventually recommended that the article be published.

Dedication

To my parents, who for some reason never once questioned whether I should be pursuing a doctorate in philosophy and never suggested that it was time to give up on my dream. I hold you responsible for instilling in me a deep, perhaps pathological concern for morality and justice.

ABSTRACT OF THE DISSERTATION

Shared Agency in the Kingdom of Ends:
Toward a Social Constitutivism

by

Zachary Charles Bachman

Doctor of Philosophy, Graduate Program in Philosophy
University of California, Riverside, 2018
Dr. Peter Graham, Chairperson

Constitutivism is the view that morality is grounded in features constitutive of our agency. This dissertation develops a version of constitutivism that grounds morality in the nature of our *social agency*. It argues that the moral law is a constitutive principle of shared agency.

Table of Contents

Preface

I believe that morality is objective. When something is wrong it is wrong, regardless of who you are, or where you come from. How can a "should" or an "ought" be objective? Answering this question is one of the tasks of a foundational theory of morality. It is what this dissertation is about.

I used to think that the only way to answer this question was by positing mind-independent, irreducibly normative properties, a "moral fabric stitched into the cosmos," if you will. I have since come to believe that this sort of an explanation is really no explanation at all. The Philosophical Project – insofar as there is one project under that heading – is to come to understand our place in the world. Positing *sui generous* metaphysical entities to explain moral objectivity is to give up on this project. It doesn't really explain anything.

This dissertation project is a kind of experiment. It started as an attempt to understand a certain explanation of morality's objectivity that I came to believe was much more promising than what the moral realist was offering. What I wanted to know was whether I could be persuaded that such a view could really work.

The view that this dissertation is about is a view that has come to be known as *constitutivism*. This is the view that morality is grounded in features constitutive of our agency. My dissertation explores a particular version of constitutivism, namely one defended by Christine Korsgaard. I don't try to show that constitutivism offers us the best of all possible explanations of morality's objectivity, or even that it is better than some other competing views. In fact, I don't even argue that the version of constitutivism I

focus on is better than other versions of constitutivism. My aim here is primarily exploratory. I wanted to figure out what the moving parts of the view are and how we can put them together to make the view work.

I am particularly interested in developing Korsgaard's view in a certain direction. I believe that morality is essentially a social phenomenon. As such, a proper account of morality's foundations should reflect this fact. Insofar as I succeed in making a contribution to the Philosophical Project, I believe that I show how we can go about constructing a *social* constitutivism, according to which morality is grounded in that aspect of us that makes us social agents. By grounding morality in our sociality, I believe that we will arrive at a richer understanding of our moral practices and come to better understand why we think morality is so important.

If the reader gets to the end of this dissertation she will see that the view I develop isn't successful in fully capturing morality's objectivity. It is important to keep in mind, however, that this doesn't mean that Social Constitutivism, as I call the view, is doomed to failure. This dissertation really just represents the initial stages of my research; this is an ongoing project. I am optimistic that, given the foundation that I have built for the view here, Social Constitutivism will eventually succeed in helping us understand the peculiar institution that we call *morality*.

Chapter 1: Introduction – What is Constitutivism?

1 Introduction: Foundations of morality

This work is an investigation into the possible foundations of morality. It is, to be more precise, an investigation into a *particular way* of working out those foundations. Perhaps the best way to begin the investigation is to say something about what a foundational investigation of morality *is*.

A foundational theory of morality provides an account of why morality has the structure and nature that it does. Of course, what this structure and nature *are* is a matter of dispute among laymen and philosophers alike. Some hold morality to be relative to time and place, while others believe it to be an eternal truth remaining fixed *regardless* of time and place. Some hold morality to be an artifact of human construction, while others believe it to be as fundamental to the nature of the cosmos as the laws of physics or mathematics. Some hold the fundamental concern of morality to be the consequences of our actions, while others believe it to be primarily concerned with the nature of our actions. A full foundational theory of morality will pick and defend a side in these and other debates and, as a component of such a defense, will provide an explanation of why morality manifests itself in the way it does.

I'm of the mind that it is a conceptual truth that morality, if it is anything at all, is a universally and categorically binding system of norms.[1] To say that morality is

[1] The remark, "if it is anything at all" is important to the claim I am making, at least as far as accurately capturing my view on the matter is concerned. It could be that I am confused about the nature of morality, that there is no universally and categorically binding system of norms that govern interpersonal conduct. If this is the case, then I'm inclined to think that morality is a *mirage*.

universal is to say that what morality requires of us is applicable to everyone regardless of time and place. To say that morality is categorically binding is to say that morality's requirements do not depend on the manifestation of certain contingent desires in order for them to bind individuals. Morality, I think, is also deontological in structure. Moral requirements concern the nature of our action, and in particular the way in which we *interact* with others. This contrasts with the consequentialist conception of morality, according to which the primary concern of morality is the maximization of states of affairs with intrinsic value.

As I see it, much of morality can be captured by Kant's formulation of the categorical imperative often called the *formula of humanity*:

> So act that you use humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means. [2]

For brevity sake, I shall call this way of thinking about morality – universality plus categoricity plus the formula of humanity – the *Kantian conception of morality*.[3]

Now, I'm not going to defend the Kantian conception as the correct conception of morality in this dissertation. Rather, I will take it as a *starting point* for my theorizing. The question I will be asking is this: what could account for a morality so understood?

One reason that some people reject the any conception of morality according to which the norms are categorically and universally binding (be it Kantian or consequentialist in nature) is that they believe we could account for such a normative

[2] Kant (2012/1785). I will discuss in quite some detail how I think we should interpret the formula of humanity in Chapter 4.

[3] In other words, according to what I am calling the Kantian conception of morality, the formula of humanity is a universally and categorically binding moral principle.

framework only with an ontologically extravagant metaphysics that includes irreducible, mind-independent normative entities to serve as "truth makers" to our moral claims. Given methodological considerations such as Occam's Razor, all else being equal I'd prefer an account of morality's foundations that is parsimonious. So, given two theories, both of which are equal in their ability to account for the central features of the Kantian conception of morality, we ought to favor that theory with fewer ontologically extravagant commitments.

The primary aim of this dissertation is to explore a *particular strategy* for providing a foundation for a Kantian conception of morality, namely one with only minimal metaphysical commitments. My present interest in this strategy is seeing how far I can develop it in a certain direction. I won't be arguing that this strategy is better than other competing strategies. Rather, this dissertation can be considered the first step to such an argument. My interest here is in understanding *how* this particular strategy works and whether it can be developed in an interesting direction. The arguments contained in this dissertation, then, are more of an exploratory exercise than a full-blooded defense of a particular view. So what we have here might be best described as an articulation of a package of ideas.

The theory that I will be exploring in this dissertation often goes by the name "constitutivism". In general, constitutivism is the view that the fundamental principles of practical reason – the fundamental principle of morality being perhaps the most important one – are grounded in features constitutive of our agency. My interest with constitutivism, at least for the purposes of this dissertation, lies, as I've already

mentioned, in developing constitutivism in a *particular direction*. I will be taking as my starting point the version of constitutivism developed in recent years by Christine Korsgaard. The view I ultimately end up with at the end of this dissertation will decidedly not be her view. But it will be *closely related to it*. Before I say more about the direction in which I will be developing constitutivism, it will be helpful to say more about what this thing I call constitutivism is.

2 What is Constitutivism?

In recent years, moral theorists working on the foundations of morality have turned to the philosophy of action.[4] These theorists – whom, following the vogue, I will be calling *constitutivists* – think that investigating the nature of our agency can provide us with a foundation to a universal and categorical morality, and do so without having to appeal to a mind independent moral fabric of the universe. The hope of constitutivists is that we can ground the norms of morality in features constitutive of our agency, in particular in those features that distinguish human action from mere behavior.

The motivation for taking the constitutive turn starts with an insight about norm-governed activities. Consider chess, which is constituted by chess' *rules*.[5] These rules set out the objective of the game, what the various pieces are, where their initial placement is

[4] Prominent philosophers who have taken this turn include Christine Korsgaard (2008, 2009), J. David Velleman (2000, 2009), Paul Katsafanas (2013), Michael Smith (2013, 2015), Luca Ferrero (2009), and Andrews Reath (2006, 2010).

[5] The chess analogy as motivation for constitutivism is, by now, a well-worn one. See, for example, Enoch (2006, 2011), Ferrero (2009), Katsafanas (2013), Silverstein (2015), Bratu and Dittmeyer (2016), Bertea (2013), Hanisch (2016).

on the board, and how these pieces ought to be moved. The rules tell me, for instance, that the knight is the piece that looks like a horse, and that its range of movement is up two and over one, or up one and over two. These rules distinguish genuine chess moves from non-chess behaviors. But in doing so, these rules also govern *my* behavior. They tell me what I *ought* to do if I want to move the knight. The constitutivist asks whether the principles of morality might be to agency what the rules of chess are to chess. Call this, the *core analogy*.

What is interesting about the core analogy, and thus what makes it worth reflecting on, is that the norms of activities like chess are not mysterious. They are grounded in the game of chess. If we can similarly ground morality in agency, we perhaps can identify a non-mysterious source of morality.

Of course, the rules of chess are neither categorical nor universal. They bind agents only when they are playing chess. In this sense, whether the norms of chess bind a particular agent would seem to depend on whether that agent has a desire to play chess. If the *core analogy* holds, how will *constitutivism* help us to explain the purported categoricity and universality of morality?

The hope that constitutivism will ground norms that are universal and categorical rests on an important *disanalogy* between chess and agency: agency, unlike chess, is *inescapable*.[6] While chess is an optional endeavor, the game of agency is not.[7] Insofar as agency is inescapable, the norms constitutive of it will (presumably) be inescapable as

[6] See Korsgaard (2009) for an articulation of this idea and Ferrerro (2009) for a defense of it.

[7] Korsgaard (2009), channeling the French existentialists, suggests that it is our "plight".

well. And if the norms of agency are inescapable, then they will (presumably) also be universal and categorical. Call this the *core disanalogy*.[8]

If the core analogy and the core disanalogy can somehow be vindicated – if we can somehow show that agency is governed by inescapable constitutive norms – and if it turns out that one of these norms is the formula of humanity, then we will have a potentially promising explanation of morality's foundations on our hands.

2.1 *Constitutive Principles, Constitutive Aims*

There are two strategies for developing constitutivism. According to the first strategy we identify the constitutive norms of an activity or practice by first identifying constitutive aims of that activity or practice. A constitutive aim of an activity or practice is an aim one must adopt in order to engage in the activity or practice in question. Perhaps Chess is such a practice, the constitutive aim of which is to checkmate your opponent.[9] The important thing about constitutive aims, so some constitutivists think, is that they give rise to practical reasons. Adopting an aim gives us a reason to do that which is conducive to aim-fulfillment. In virtue of adopting the aim of check-mating my opponent, I have a reason to, say, move my knight up one another and over two as this will increase my ability to force you into checkmate.

---

[8] Whether or not securing the disanalogy between games and agency in this way successfully grounds categorical and universal norms has been the focus of much debate concerning the success of constitutivism. This is an important question, but not one that I will be investigating in this dissertation.

[9] Katsafanas (2013) suggests this.

Theorists who adopt the constitutive aim strategy try to identify a constitutive aim of agency. The hope is that this constitutive aim – whatever it turns out to be – can serve as the source of our moral reasons, and thus serve as morality's foundation.

The second strategy for developing constitutivism is what I call the *constitutive principle strategy*. A constitutive principle is principle that is descriptive of and normative for engagement in an activity or practice. The simplest kinds of constitutive principles are rules of games. In order to count as playing a particular game, one must comply with the rules of that game. The rules of chess, for example, tell you to move the knight up one, and over two, or up two and over one. The rules of baseball tell you that if you want to throw a pitch, you have to start with your foot touching the "rubber" on the mound. Failure to comply with a constitutive principle entails failure to engage in the activity the principle is constitutive of. But constitutive principles also tell you what you *should* do in order to engage in a certain activity or practice.

Theorists who adopt the constitutive principle strategy hope to show that the moral law is in some way a constitutive principle of agency—it describes, in some way, how to engage in the activity of agency. In this dissertation I will be adopting the constitutive principle strategy, not necessarily because it is a better alternative to the constitutive aim strategy (though, I think it is), but more so because it is the road less-traveled.

3 Varieties of Constitutivism

There are three main players in the constitutivism literature: Christine Korsgaard, J. David Velleman, and Paul Katsafanas.[10] This dissertation largely focuses on the constitutivism of Christine Korsgaard. I do this for a few reasons. As we will see shortly, Korsgaard's constitutivism is the most ambitious; unlike Velleman and Katsafanas who derive something *approaching* morality out of what is constitutive of agency, Korsgaard believes that she gets all of enlightenment morality.[11] The second reason I focus on Korsgaard's constitutivism is that it is widely misunderstood. One aim of this dissertation is to show how we can get her right. Finally, I focus on Korsgaard because this project started as an attempt to try to understand *how* her view works. If this dissertation makes no other contribution, hopefully it will at least make some progress in our collective understanding of an important contemporary philosopher. Along the way I will disagree with some of Korsgaard's claims and point out places where she makes argumentative mistakes. But the view I end up endorsing in the end, though certainly not *Korsgaard's* view, is certainly Korsgaardian in *spirit*.

In section 4 I will be offering an interpretation of Korsgaard's version of constitutivism. But before I do this, I should say something about Velleman's and Katsafanas' versions of constitutivism, both of which are constitutive aim views. I will be

[10] These, of course, are by no means the only constitutivists around. An increasing number of philosophers have been taking the constitutivist turn in recent years, including Caroline T. Arruda, Luca Ferrero, Jeremy Fix, Kathryn Lindeman, Andrews Reath, Michael Smith, and Kenneth Walden, among others. I state that Velleman, Korsgaard, and Katsafanas are the three main players because they have all written monographs on the subject and are treated as such in much of the secondary literature on constitutivism.

[11] Whether she succeeds in this is another matter.

brief here, since most of what I say has no bearing on the rest of the dissertation. I spend space on these accounts so that the reader can get a sense of the other flavors of constitutivism that are on offer.

3.1 David Velleman's Self-Understanding Constitutivism

The first step for constitutivists who adopt the constitutive aim strategy is to identify a constitutive aim of action or agency. [12] A constitutive aim of an activity or state, recall, is typically defined as that toward which every instance of that kind of activity or state aims.[13] Constitutive aims are helpful because they are generally thought to give rise to standards of success. Consider the mental state of belief, which is widely considered to have truth as its constitutive aim. To say that truth is belief's constitutive aim is to say that every instance of belief aims at truth, and that, because of this, truth is the standard of success for belief.

Velleman's constitutivism begins by reflecting on the case of belief, in particular on the fact that we can rather easily generate an account of epistemic norms from the fact that belief has a constitutive aim. On this account, the principles of epistemic reasoning are those principles which by following we will likely arrive at the truth. The normativity

[12] In this dissertation I will largely be using 'action' and 'agency' interchangeably, using whichever word sounds better to my ear in the sentence in question. Strictly speaking, these words are not synonymous, at least for most theorists of action. The term 'agency' is typically used to refer to a capacity, the term 'action' is used to refer to an exercise of that capacity. If my looseness here bristles the reader, feel free to replace as you see fit.

[13] I say 'activity or state' because, as we will see shortly, many think that mental states can have constitutive aims.

of these principles is grounded in the constitutive aim of the state they govern. Velleman's insight was that the structure of epistemic normativity so understood might be applicable to other kinds of normativity, in particular normativity in the practical domain. Might action, like belief, have a constitutive aim?[14]

Velleman, of course, thought the answer to this question was yes. For Velleman, the constitutive aim of action is intelligibility, or self-understanding as he sometimes puts it.[15] Over the course of his career, Velleman unpacks this idea in different ways. In his (1989), he attributes to all reflective agents a desire never to be caught unaware. This constitutive desire gives us reason to choose those actions which will makes sense for us to choose given the particular motivational set that we have.[16]

In later work, Velleman draws an analogy between acting and improvisation. What we do when we set out to perform an action is similar to what an actor does on stage when she is performing improv. An improvisational actor is not given a script, but rather a scene and a character, and is told to interact with the other actors who similarly were given just a character but no script. What does the actor do? She tries to act in a way that makes sense, given her character and the scene she finds herself in. Much like the improvisational actor, we, every day agents that we are, must also figure out what to do without the use of a script. How do we go about doing this? According to Velleman,

---

[14] Velleman relates action to belief in this way in the introduction to (2000). See in particular pp. 15-31.

[15] It should be noted – and this is pretty common in the literature on constitutivism – that when Velleman speaks of "action" or "agency", he typically means the action or agency of reflective creatures, in particular humans. This isn't problematic in itself as long we track what he means by the term.

[16] An important component of this (1989) account is that, for Velleman, intentions are self-referential beliefs akin to predictions.

much like the improvisational actor, we try to choose actions that make sense, given who are and where we find ourselves. If intelligibility, so understood, is the constitutive aim of agency, then what we have most reason to do is whatever will help us make the most sense of ourselves.

How does intelligibility qua constitutive aim of agency provide us with a foundation for morality? Many of the things we do take place in the social world, in the sense that our actions are socially defined. To borrow Velleman's example, when we go out to eat at a restaurant there are a series of moves that we make (approach the maître d', wait to be seated, order from the menu, etc.) which are sanctioned by conventional scenarios. These conventional scenarios are important, for they make it possible for us to interact in ways that we otherwise couldn't. The restaurant scenario is what makes *eating out* possible. The going-to-a-movie scenario makes *watching a movie in public* possible. The student-at-school scenario makes *education* possible. Without these or other conventional scenarios, social life would involve a much higher degree of chaos.

From a certain perspective, the content of these scenarios is arbitrary. They could be almost anything. What matters is that there be some way for us to converge on a scenario so that we can coordinate our behavior with each other. That being said, Velleman thinks that, given our sociality, we are strongly inclined to develop a repertoire of such scenarios that we can share. Because we are rational, we will be inclined, over time, to develop scenarios that are more coherent and consistent with how we understand ourselves. As our social systems grow, and the social webs we weave become more and more interconnected, we will be increasingly inclined to treat others as fellow

collaborators, rather than hostile forces. These two forces – rationality and inclusion of others – will lead us to eliminate arbitrary distinction that preclude others from the social scripts we engage in. Velleman, reflecting on the process he describes, suggests that this is what we often have in mind when we talk about "moral progress". In this sense, practical reasoning has a *pro-moral tendency.*

Now, it is important to note that for Velleman, there is no *guarantee* that agents constructed in the way that we are will, over time, develop morality. Nor, does Velleman seem to think that it is the only way for beings like ourselves to solve the coordination problem of living together. Practical reasoning has a pro-moral *tendency*, not a pro-moral *necessity*.

3.2 Paul Katsafanas' Nietzchean Constitutivism

Most versions of constitutivism are Kantian in spirit. Katsafanas breaks from this strategy by developing a version of constitutivism that is rooted Nietzsche's normative theory.

There are two parts of Katsafanas' view that are worth talking about, one concerning the structure of the view, the other concerning the content of the view. We'll start with the structural point. The starting point for Katsafanas' positive account is a problem that he identifies in Velleman's view. Constitutive aims, Katsafanas argues, come in two flavors: simple aims and differentially realizable aims. Katsafanas defines a simple aim as "an aim that a particular action will either fulfill or fail to fulfill" (2013, p. 75); he defines a differentially realizable aim as "an aim that can be fulfilled to different

degrees". Compare the following two aims that an individual may have. Young philosopher A has an aim to get a job after graduate school. Young philosopher B, on the other hand, has an aim of getting a *good* job after graduate school. According to Katsafanas, philosopher A's aim is a *simple* constitutive aim. Either philosopher A gets a job, or she doesn't. If she does, then she has fulfilled her aim. If she doesn't then she will have failed to do so. Philosopher B, on the other hand, has a *differentially realizable aim*, for his aim can be satisfied to a greater or lesser degree (depending on how good the job is). Imagine that for philosopher B, a good job is a research-oriented position. B got an interview at three different schools, one a large state school with a heavy teaching load and light research requirements, and no graduate program; a second at a large state school with a significant but not heavy teaching load, moderate research requirements, and a terminal master's program; and a third at an R1 institution with a PhD program, a light teaching load, and significant research requirements. Philosopher A's aim (get a job) would be satisfied equally where she to get an offer from any of these schools. Philosopher B's aim (get a *good* job), however, would be satisfied to a greater extent where he to get an offer from the R1 institution than it would were he to get an offer from either of the other two schools.

Now, Katsafanas thinks that in order for a constitutive aim theory to make sense of practical reasoning the aim constitutive of agency that the theory identifies must be differentially realizable. There are two aspects to an account of practical reasoning: an account of what we have reason to do, and an account of what we have *most* reason to do. If a constitutive aim is *simple* rather than *differentially realizable* it will lead to a collapse

in the distinction between what we have reason to do and what we have most reason to do. Katsafanas argues that intelligibility is a simple aim. If options A and B are to some extent intelligible options for me, then, as far as the constitutive aim of action is concerned, it doesn't matter which of the two options I should choose. But this is not how practical reasoning works. Though there certainly are times where we are presented with two options which we have equal reason to pursue, practical reasoning does not treat two options with a *differing weight* of reasons supporting them as equally good options.

Katsafanas argues that the problem with Velleman's version of constitutivism is that the constitutive aim it identifies is a simple one. If two actions both contribute to some amount of self-understanding or intelligibility, then as far as action's constitutive aim is concerned, both options are equally choiceworthy, even if one option would lead to a *greater* self-understanding.[17]

What Katsafanas sets as his task, then, is to identify a constitutive aim of action that is differentially realizable. To do this, Katsafanas turns to Nietzsche for inspiration.

Katsafanas identifies two distinct constitutive aims of action. The first of these constitutive aims is what Katsafanas calls *agential activity*. To aim at agential activity is to aim at what Katsafanas calls *equilibrium*. In A-ing, an agent is in equilibrium if the agent "approves of her A-ing" and would continue doing so were she to come to understand the etiology of why she has come to perform A (138). An agent is in

[17] Katsafanas notes that Velleman *thinks* his account has the resources to capture how practical reasoning works in these cases, but Katsafanas argues that he is mistaken about this. I will not be concerned with who is the better interpreter of Velleman on this point, Katsafanas or Velleman himself. That being said, I'm inclined to think that intelligibility and self-understanding are differentially realizable.

disequilibrium if the agent would not continue to approve of her A-ing were she to come to understand the etiology of her action.

Katsafanas identifies equilibrium with activity in action because he thinks it better captures the idea of autonomy in action than do the more standard accounts that require deliberation and reflection. The intuitive idea behind Katsafanas' equilibrium requirement on agential activity is that those actions which we would endorse upon coming to better understand their historical roots are more reflective of who we are as agents than those actions which we would not so endorse.[18] Equilibrium *qua* constitutive aim tells us that we have reason to choose those actions that we would endorse upon discovering their etiology.

To engage in reflective agency, argues Katsafanas, is to aim at determining one's action via choice. One determines one's action by choice when one achieves equilibrium in action. It follows that the constitutive aim of reflective agency is agential activity, the correct analysis of which, according to Katsafanas, is equilibrium in action.

Katsafanas recognizes that the first constitutive aim does not result in a robust account of practical reasoning. In particular, the view can lead to wildly different conclusions depending on what desires and values one happens to have. What Katsafanas needs is a way to constrain the normative options, to rule out certain courses of action, even if they'd be endorsed by an agent upon reflection. This is where Katsafanas introduces a second constitutive aim: the will-to-power.

---

[18] One might think that the view Katsafanas is suggesting is no different from the view that Korsgaard defends in *Sources of Normativity*. While they may be close, they are not the same. Korsgaard's idea of reflective endorsement goes through the concept of a practical identity – a description under which you value yourself – and is not necessarily concerned with genealogical critique.

Will-to-power, as Katsafanas understands it, is "a formal relation, which describes the structure of willing" (162). To will-to-power is to seek and overcome resistance in the pursuit of one's end. This describes, not an end to seek for its own sake, so much as a *way* or *method* of willing. We satisfy this aim when seek ends that are difficult to pursue (such as artistic endeavors) or pursue in difficult ways our ends that we already have.

Setting aside the question of the plausibility of will-to-power as a constitutive aim of agency, let's turn to consider how these two constitutive aims interact. The constitutive aim of will-to-power provides a kind of normative constraint on satisfying the constitutive aim of seeking equilibrium in action. When we combine the two constitutive aims together, we arrive at the conclusion that, among those actions that we would endorse upon learning about their etiology, we have more reason to pursue those actions that offer greater resistance than those actions that don't.

Is Katsafanas able to generate morality out of these dual constitutive aims? He argues that what we arrive at is something approaching the conclusions that Nietzsche arrived at: we should reevaluate our values in accordance to how well they satisfy our aim of will-to-power. Of course, Nietzsche's moral theory is decidedly *not* the Kantian conception of morality I set out to defend in this dissertation.

3.3 Concluding Thoughts on the Constitutive Aim Strategy

The constitutive aim strategy, on its face, has a lot going for it. I will not provide any knockdown arguments against it. My interest in this dissertation is not to show that the strategy I pursue is the best or correct strategy. Again, my aim is an experimental one:

how far can I take a constitutive principle approach? But before we leave the constitutive aim strategy behind for good, it will perhaps be helpful to make a few general remarks about it.

The first remark I'd like to make concerns the fact that the two main constitutive aim theories make use of highly controversial accounts of agency. That this would be the case should not be all that surprising. Intuitively, if we hope to derive a robust system of norms from a constitutive aim of agency, this constitutive aim will have to itself be robust. The problem here is that the more robust the aim we identify, the more controversial the account of agency we end up with. A more minimal aim will be less controversial but will be much less likely to provide us with what we need to derive morality.[19]

Second, the constitutive aim strategy is not as straightforward as it first appears. When Velleman appeals to the case of belief he writes that truth is a standard of correctness for belief *because* belief descriptively aims at truth. Perhaps this is so, but it belies the recent history of the idea. Generally, Bernard Williams (1973/1995) is credited with the idea that truth is belief's constitutive aim. But in his famous paper, "Deciding to Believe," Williams offers three reasons for attributing to belief the constitutive aim of truth. The first is that "truth and falsehood are a dimension of an assessment of beliefs" (p. 137). The second is that "to believe that p is to believe that p is true" (ibid). The third is that the utterance 'I believe that P' "carries, in general, a claim that p is true" (ibid). In

[19] This point is made in Setiya (2007). He calls the view he criticizes "ethical rationalism". This is just an idiosyncratic way of referring to constitutivism.

the first of these, it would seem that Williams is pointing out that truth is the standard of correctness for belief. In the second of these, it would seem that he is pointing out what Velleman would call belief "descriptively" aiming at truth. What is the relationship between these two? Williams is not clear. But some philosophers, e.g. Wedgewood (2002) and Engel (2013), have suggested the relationship is the opposite of what Velleman identifies. Belief descriptively aims at truth *because* it is the standard of success, not vice versa.

I'm not sure which of these ways of unpacking the concept constitutive aim is the correct one. I raise this worry only to point out that the concept is not a trouble-free one for the constitutivist.

Neither of these points are meant as knock-down arguments against the constitutive aim strategy. In fact, neither of them is meant to be an argument against them *per se*. Whatever the case, there is another potential way to make constitutivism work. This is the constitutive principle strategy that I pursue in this dissertation. And it is one that Korgaard pursues in her work. Let's turn to her view next.

4. An Interpretation of Korsgaard's Constitutivism[20]

4.1 Nuts and Bolts I: The Structure of Action

Korsgaard's constitutivism combines elements of Neo-Kantian and Neo-Aristotelian thought. Neo-Kantians take the principles of practical reason to be grounded in the

[20] Many of the details in this section will become relevant in Chapter 3 when I concern myself with Korsgaard's bad action problem.

structure of our *will*.[21] Aristotelians take the principles of practical reason to be grounded in the *teleological structure* of the kind of beings that we are.[22] Korsgaard combines these two theses: the principles of practical reason are grounded in the *teleological structure of our will*. In particular, she argues that "the principles of practical reason serve to unify and constitute us as agents," much in the way that teleological organization unifies mere heaps into artifacts (2009, p. 27). In order to engage with Korsgaard's view, we must understand what she means by this.

Our story begins with a familiar Aristotelian yarn concerning the tripartite structure of artifacts. An artifact is composed of *matter* that is arranged according to a certain *form,* which enables the object to fulfill its *ergon* or "purpose, function, or characteristic activity" (2009, p. 27). To illustrate, Korsgaard has us consider a house: its matter is the stuff (e.g. bricks) out of which it is constructed. This matter is arranged in a certain way—this is its form—which enables the structure to *serve as shelter*, which is its ergon.

A teleologically organized object exhibits an instrumental structure in that it is organized around achieving an end, namely the object's ergon. Teleological organization is significant because it "supports normative judgments" about objects (2009, p. 28). A house is good *as a house* in virtue of the way it is organized toward achieving the house's ergon. Normative judgments grounded in teleological organization "meet skeptical challenges to their authority with ease" (2009, p. 29). It makes little sense to ask why I

[21] By 'neo-Kantian' I mean such figures as Christine Korsgaard, J. David Velleman, Onora O'Neil, and Andrews Reath.

[22] See, for instance Foot (2001) and Thompson (2008).

should arrange matter in a way conducive to serving as shelter if what I have committed to building is a *house*.

The next step in our story involves applying this picture to the case of agency. It is a bit of a process, so please bear with me.

Korsgaard adopts a broadly Aristotelian understanding of nature, where living things are understood to exhibit the tripartite structure of artifacts. On this view, a living thing "is a thing so designed as to maintain and reproduce its form. It has what we might call a self-maintaining form" (2009, p. 35). Maintaining and reproducing its form is a living thing's *ergon*, which is carried out by, for example, eating, running from predators, and mating. The *matter* of a living thing is the various things we find when we dissect it, such as organs and tissues.

While at a certain level of description all living things have the same *ergon* – namely self-constitution – they do not all have the same form, as every species is organized differently. Following Aristotle, Korsgaard notes that all living things are vegetative, meaning they are organized so as to grow and reproduce. Animals, in addition, are organized to maintain their form through perception and action. Human animals have yet a further layer of organization: we are organized to maintain our form through *rational activity*.

Korsgaard thinks that action, itself, can also be understood in terms of this artifactual model. This means that action also has an ergon, a form, and is composed of matter.

For Korsgaard, our actions are comprised of act-types coupled with ends. *Running* is an act type. *To evade a predator* and t*o catch prey* are two ends running can serve. Acts and ends are tied together by a *principle*, which is the *form* of an action. When we articulate the principle behind an action, we articulate why the action—the act for the sake of the end—is choiceworthy. If the action I choose is *running in order to evade a predator*, the principle behind my action might be articulated as so: *running is the thing to do when evading a predator*. A principle is the *form* of an action because principles articulate relations of rational support that hold between acts and ends. In doing this, they explain why an act makes sense in light of an end.[23] In this sense principles *organize* actions.

We are now in a position to understand what Korsgaard has in mind when she says that "the principles of practical reason serve to unify and constitute us as agents". According to Korsgaard, animals, like artifacts, are teleologically organized. An animal's ergon is to be and continue being the kind of animal that it is. An animal's form is its species-specific organization, which will include, among other things, the capacity to act. Action *itself* is also teleologically organized. The ergon of action is whatever action's role happens to be in the self-maintenance – or more generally the life – of the organism. The form of action is the principle that organizes action. The important take-home point here is that agency is organized by principles. These principles help serve an animal's immediate ends. But in serving such ends they also help serve the end all animals have as self-maintaining organisms. We act, among other things, in order to maintain ourselves.

[23] For non-human animals instinct plays the role of a principle.

But *in order* to act, we need to be unified and constituted as agents, for, according to Korsgaard, action is attributable to an agent as a whole, and if we are not unified, our behaviors would not be attributable to us as a whole but rather to the sub-agential motivations moving through us (2009, pp. 18-19, 125-126).

4.2 Nuts and Bolts II: Self Constitution as a Function of Action

In the previous section it was suggested that the ergon, or function, of action will be that role that action plays in the self-maintenance of an animal. Intuitively, acting enables an animal to better perform the four Fs: flee, fight, feed, and reproduce. For example, acting enables an animal to run from predators, fend off an attack, forage for food, and find a mate, things sessile organisms cannot do. In paradigmatic cases, an individual effects a change in the world (e.g., knocking an orange off of a tree), and if all goes according to "plan," this change will contribute to the animal's maintaining its form.[24]

We earlier characterized an action as an act-end pair organized by a principle. We can now see that the end of an action is a kind of commitment on the part of the agent to effecting a change in the environment. This commitment to efficacy "is built right into the structure of agency," for behaving without such a commitment is to just flail about, but "to flail is not to act" (2009, p. 88). For Korsgaard, this is true of lower animals as

[24] Of course, not every action contributes to the self-maintenance of the acting animal. But this is no more a problem for the current claim than is the fact that every wing-enabled flight does not contribute to a bird's self-maintenance. The claim on the table is that action, *in general*, contributes to the self-maintenance of the organism, not that every action does. One can think of agency as a trait that some organisms have but others (such as plants) lack. Presumably, organisms developed agential capacities due to the role they play in the pursuit of the four Fs.

much as it is of humans. When, for example, a dog digs for a bone, it commits to effecting a change in the world, namely to digging up the earth that the bone lies under. [25]

So acting requires a commitment to efficacy. But efficacy, it turns out, requires one's agency to be *unified* or *constituted*. Reflective creatures are capable of questioning those things which we are inclined to do. When entering the kitchen for breakfast, for example, we may be tasked with a difficult choice: donuts or oatmeal? *The donut looks tasty, but unhealthy. Oatmeal is healthy, but bland. Which should I eat for breakfast*? This gap between our experience of an incentive and our acting on that incentive must be closed in order to pursue an end. If the gap is left open, we will be in disunity -- a pile of drives competing for dominance -- perhaps until one of them moves us to act heteronomously. Efficacy in the pursuit of our ends requires that we constitute or unify ourselves by bridging the gap created by reflection.

Korsgaard argues that we bridge this gap by adopting a principle that can adjudicate between competing inclinations. I might act on the principle of desire satisfaction and choose to eat the donut. Or, recognizing that donuts are toxic to my health, I could respect my humanity and choose to eat the oatmeal.

Importantly, not all principles are equally up to the task of bridging the gap created by reflection. In having a law-like structure, principles have future implications for the agent acting on them. Were I to act on the principle of desire-satisfaction in my

[25] This is not to say that the dog *represents* itself as so committed. A commitment may be a first-order, non-conceptual representational state, akin to visual perception in its structure. Just as an animal can objectively represent its external environment without being able to represent itself (see Burge 2010), so could an animal have a commitment that doesn't involve meta-cognition. On such an account a commitment may be an aspect of an intention-like state. See chapter 2 for a discussion of this point.

decision concerning what to eat, I would be willing that whenever I am hungry I should eat whatever I most desire. But my appetites are often shifting. For dinner, I may begin to prepare pasta when I notice a candy bar on the counter. This may lead me to want to eat candy. Since I am acting on the principle of desire-satisfaction, my dinner plans will have shifted with the shift in desire. However, as I open the wrapper, I may notice that my cookie jar still has cookies in it. This may lead me to desire a cookie. Insofar as my appetites are a forever shifting landscape, willing that I eat whatever I most desire might result in my never eating a bite.[26] Of course, if my appetites remain stable, none of this may happen; I'll just continue making pasta. But this is something I don't have control over. So the principle of desire satisfaction is conducive to efficacy and unification only when contingent conditions I have no control over obtain.

For Korsgaard, the only principle truly up to the task of bridging this gap is the moral law, as that is the only principle that does not require the obtaining of such contingencies (2009, p. 180). The explanation Korsgaard provides as to why is a bit obscure, but the story goes something like the following. Efficacy in the pursuit of our ends is a *diachronic* affair: the pursuit of an end is an ongoing exercise realized by performing many actions over time. For Korsgaard, who seems to hold a Parfitian view of personal identity, my relationship to my future self is structurally similar to my synchronic relationship with another person.[27] This suggests that in order to execute a

[26] This is a version of Korsgaard's familiar example of Jeremy, the young college student who never accomplishes a thing. See (2008, p. 127) and (2009, p. 169).

[27] See Parfit (1984).

diachronic action, I must coordinate with my future self, similar to how sharing my agency with another requires coordinating with that person. In other words, diachronic agency is agency I share with my future self. Now, in order to share my agency with another person (or my future self), I and the other person (or my future self) must "unify our wills"; we must "deliberate together" in order to arrive at a "shared decision" (2009, p. 190). In order to deliberate *together*, and not *against* one another, we must treat each other as ends, not mere means, and this involves acting only on principles that we can *will* as universal laws, that is, laws that everyone can act on. If I act on a principle that you refuse to act on, then you will reject the principle, and we will fail to act together. So in order for us to come together as an efficacious unified agent, we must act on universal principles. And this, according to Korsgaard, is just what the moral law directs us to do. But now, insofar as action is "interacting with yourself", efficacious diachronic agency requires acting on universal principles as well, for "the law that you make for yourself now must be one you can will to act on again *later*". If it's not, then I won't be efficacious in the pursuit of my diachronic ends. So the moral law is the only principle truly up to the task of unifying our agency across time.[28]

In acting we aim to be efficacious. Efficacy is achieved only to the extent that we are unified agents, while the extent to which we are unified depends on the principles we adopt *in acting.* So, we act not only to be efficacious, but also *to constitute ourselves*. Action, then, has a dual function: efficacy and self-constitution.[29]

---

[28] In chapters 4 and 5 I will be discussing in more detail the controversial claims made in this paragraph.

[29] Sometimes Korsgaard states that, rather than self-constitution, the second function of action is "to render [the agent] autonomous" (2009, p. 83). But by this she means much the same thing as self-constitution. On

5. Developing Social Constitutivism: An Outline of the Dissertation to Follow

From the above discussion, it will be noticed that there are two aspects to Korsgaard's self-constitution model that correspond to actions' two functions. We can call these the efficacy side of action and the self-constitution side of action.

Korsgaard, as we saw, thinks these are intimately intertwined. Efficacy requires self-constitution and self-constitution requires efficacy. However, I'm inclined to think that they are so intertwined because her starting point is one where agency is a heap that must be put together. Efficacy requires self-constitution only because agency not constituted cannot be efficacious. But what if agency is prepackaged as fully constituted? What if the self does not need to be unified? Does this spell the end for Korsgaard's constitutivism? Does it live and die on her controversial theory of the self? One of the underlying strands of this dissertation is that the answer to this question is, perhaps surprisingly, no. What makes this surprising is that most commentators focus on Korsgaard's theory of the self as the fundamental contribution she makes in *Self-Constitution.*[30] It's beyond dispute that this *is* one of her many important contributions. My suggestion is that, interesting as the thesis is, it is not essential to making a view like hers work. I will try to show that the important components of Korsgaard's view – these are the components responsible for grounding the normativity of the moral law – rest on the efficacy side of action. In particular, I will be endorsing the idea that we are

---

her view, one becomes autonomous by constituting oneself. In fact, sometimes she explicitly says that the function of action is self-constitution (e.g., 2009, p. xii). I opt for this formulation as it more straightforwardly expresses Korsgaard's view.

[30] The title is *Self-Constitution*, after all!

inescapably committed to efficacy in the pursuit of our ends. But I will be developing this thought in a particular direction. I will suggest, in chapter 2, that an important component of practical reasoning involves adopting new commitments and fitting them into a coherent web with our beliefs and other, prior commitments. I will argue that this view of practical reasoning can explain why constitutive principles are not merely descriptive, but also normative. I call this view the Commitment View.

One of the common criticisms of Korsgaard's view is that it suffers from an ailment that commonly affects constitutivists, namely the Bad Action Problem. A version of constitutivism suffers from the bad action problem if the standards it identifies that distinguish action from mere behavior are the same as the standards it posits for good action. When this is the case, such views have the unwanted implication that all actions are good and no actions are bad. I will show, in chapter 3, that with slight tweaks to her view, we can solve the bad action problem for Korsgaard. Working through the bad action problem will not just be worth doing for its own sake, but will also provide us with some insight into how some of the moving parts of Korsgaard's view work together. From this understanding we can construct better versions of constitutivism.

Another component of her view that I will be endorsing is the idea the idea that the moral law is a constitutive principle of shared agency. That this is Korsgaard's view will be surprising to a lot of philosophers. The common understanding of Korsgaard's view is that she holds the moral law to be constitutive of *self-constituted agency*. In Chapter 4 we will learn why this common understanding doesn't accurately represent Korsgaard's overall view. There I will be arguing, first, that Korsgaard's defense of the

thesis that the moral law is a constitutive principle of shared agency fails, but, second, that we can defend the thesis in a different way. So here is another instance where I borrow from Korsgaard while also veering from her.

I veer yet further from Korsgaard in Chapter 5, where I reject her way of capturing the categorical and universal normativity of morality. Unfortunately, I don't yet have an account to replace it with. In place of such an account I offer a few suggestions as to how the social constitutivist can move forward.

I call the view I develop in this dissertation *Social Constitutivism*. To adopt this view is to adopt a package of views: the Kantian conception of morality, the constitutive principle strategy for developing constitutivism, the Commitment View of practical reasoning, and the thesis that the moral law is a constitutive principle of shared agency.[31]

Ultimately, it might turn out that Social Constitutivism fails. May aim in writing this dissertation was to see what would be involved in developing the view. Can the view be taken further than I took it here? I think the answer is yes, but this might be blind optimism. I will let the reader judge this for herself.

[31] As we will see in chapter 5 I am not the only constitutivist who now uses the moniker "social constitutivism".

Chapter 2: The Commitment View

1. Introduction

Our goal is an anti-realist foundation of morality that is constitutivist in structure. The constitutivist theory that I will be developing in the chapters that follow grows out of an account of practical reasoning which I call the Commitment View. The idea behind this commitment-based account is simple. Some activities are governed by constitutive principles. Those constitutive principles become normative for us – i.e., they govern our behavior – when we commit to engaging in those activities. This chapter provides a sketch of the role that commitments play in our practical reasoning. This chapter does not provide a complete theory of practical reasoning or the role that commitments play therein. It is a sketch of a small *part* of practical reasoning. My hope is that what I present here will be an intuitive picture of practical reasoning that is compatible with a wide variety of theories.

The Commitment View is presented here really as a means to an end. The main aim of this chapter is to do two things. First, I aim to explain why constitutive principles are normative. A common objection to constitutivism is that constitutive principles, in being descriptive, cannot serve as normative principles, for constitutive principles, in themselves, provide us with no reason to comply with the behavior they describe. Thus, so the objection goes, the moral law cannot be a constitutive principle. I respond to this objection in section 4.

The second main aim of this chapter involves showing that the instrumental principle is a constitutive principle of action and explaining how the commitment view can explain the principle's normativity. This argument will have the controversial implication that every action complies with the instrumental principle, which on its face would seem to violate Korsgaard's so-called "error constraint", according to which a principle can be normative for us only if it is possible for us to violate it. I argue that while the instrumental principle does not satisfy Korsgaard's error constraint, I show that there is a different error constraint we can appeal to which leaves room for the normativity of the instrumental principle. I defend these claims in section 5.

Sections 2 and 3 provide a sketch of the commitment view that I will be applying in sections 4 and 5. This view will return frequently in the dissertation. I end this chapter, in section 6, by explaining the role that the Commitment View will play in the overall argument of the dissertation.

2. Three Philosophers on 'Commitment'

What is a commitment? The term is thrown around a lot in philosophy, often unreflectively so. Philosophers often speak of their theoretical commitments, or of what a view commits one to. We often speak of the commitments that structure a life, making a commitment to a person (I will *commit* to my spouse), a moral ideal (I commit to going vegan … after I eat this pint of ice cream), a new exercise or work regime (I will commit to working out three times a week and writing 1 hour every morning), and a new job (we need you to commit to being here for three years). We use the word 'commitment' to

mean many different things. Sometimes, these different uses are connected, sometimes they aren't. My aim in this section is to identify how three philosophers use 'commitment' in their views. I will be drawing on pieces of each of their views when I construct the Commitment View in Section 3.

2.1 Shpall: Commitments are Normative Relations

I began thinking that commitments might be able to solve some of my problems with constitutivism when I read Sam Shpall's recent papers "Moral and Rational Commitment" and "Wide and Narrow Scope". In these papers, Shpall makes use of the notion to solve some puzzle about promises and rational requirements. My interest in Shpall's work is not in how he solved these puzzles, but rather in how he identified and distinguished three different kinds of commitments: rational, moral, and volitional commitments. Let's look at how he distinguishes these three commitments:

> **Rational Commitment:** "To be rationally committed to having A is to be such that you must be irrational if you fail to have A, assuming no changes in your other attitudes." (2014, p. 148)
>
> **Moral Commitment:** "the commitment you take on by making a promise" (2014, p. 146)
>
> **Volitional Commitment** "The commitment that's constituted by your dedication to some person, object, or goal." (2013, pp. 727-728)

Let's consider some examples. Hunter is **rationally committed** to believing that the Padres won't win the World Series if he believes the following propositions: *Whoever wins the World Series will come from the Eastern Division of the American League* and *the Padres are not in the Eastern Division of the American League*. Will has made a

moral commitment to Dave if Will has promised to help Dave move to Los Angeles. Sara has made a **volitional commitment** if she has made a decision to go to law school

According to Shpall, commitments are normative relations that are grounded in, but not reducible to, the attitudes of the agents whose commitments they are. That I am rationally committed to believing q is grounded in my belief that p and my belief that p implies q. But the rational commitment itself is not an attitude that I have; it is a normative relationship between the attitudes I have and the attitudes I ought to have. That I am morally committed to ϕ-ing is grounded in, but not reducible to, the promise I made to my partner. The moral commitment is not the promise itself, but rather the normative relationship between the promise I make and the intention I ought to form in order to fulfill that promise. According to Shpall, "The commitments themselves are normative, in the sense that they put genuine pressure on the committed agent to form the attitude to which he's committed" (2014, p. 149).

There are a couple of questions I have about Shpall's view that I am not able to answer. First, what does Shpall mean by "genuine pressure". Is this the "pressure" of the metaphorical kind, such as the "normative pressure" often invoked by moral philosophers? Or is the pressure of the psychological kind, perhaps in the family of motivation?

The second lingering question I have about Shpall's view concerns the third kind of commitment listed above. Is a volitional commitment also a normative relation? If so, what are the two relata in the relationship? Perhaps the normative relationship concerns

the relationship between means and end. Unfortunately, Shpall says very little about this kind of commitment. His work primarily focuses on the other two.

2.2 Gilbert: Commitments of the Will

Margaret Gilbert is famous for arguing that a lot of social phenomena – shared actions, shared beliefs, promises, agreements, and more – are made possible by what she calls *joint commitments*. Chapter 4 will spend a lot of space discussing Gilbert's notion of joint commitment. Here I will be interested in a different kind of commitment that she writes about, namely what she calls "commitments of the will".[1]

As Gilbert understands the concept, a commitment of the will is what is involved in "the personal decision of an individual human being" (2007, p. 262). Gilbert notes that when we make a decision, we thereby commit ourselves to that course of action. She adds that "if anything is an exercise of a person's will, his making a decision is" and that it is for this reason that she calls this variety of commitments "a commitment 'of the will'" (ibid).

Gilbert identifies two features of this kind of commitment. First, as has been mentioned, it is a "creature of the will" (ibid, p. 263). By this, I take it, she means that it is a product of volition. Second, these commitments "bind" the will in two different ways. According to Gilbert, the first way in which commitments bind the will is the sense in which "one who makes [a] decision has sufficient reason to act in accordance with it" (ibid, p. 263). Here, Gilbert is not identifying the kind of reason that one takes up in

[1] See Gilbert (2007).

practical reason when deliberating about what to do. In other words, Gilbert is not saying that one can give oneself a reason to act simply by deciding to so act. Rather, Gilbert is suggesting that there is a sense in which reason requires one to do what one has decided to do.

The second way in which these commitments bind the will concerns what she calls persistence: "Given one's decision, one will continue to be bound in the first sense [i.e. have reason to do what one has decided to do] unless and until one has carried out the decisions – unless or until one rescinds it." (ibid). Gilbert adds that intentions, which are also "creatures of the will", are bound to the will only in the first sense, not in the second sense:

> One *can* decide not to do what one finds oneself with an intention to do and hence deliberately repudiate it. It will then cease to bind in the first sense at issue. It may cease so to bind, however, without any such repudiation. It can simply go out of existence or be replaced by a contrary intention. (ibid)

I think the way in which Gilbert draws the distinction here between intention and decision is wrong. I believe that both decisions and intentions bind the will in this second way. But not much hangs on my disagreement with Gilbert on this point.

2.3 Bratman: Intentions as Commitments

In his important work, *Intentions, Plans, and Practical Reason*, Michael Bratman suggests that an intention "involves a characteristic kind of commitment" (1987, p. 15). He identifies two dimensions of commitment, what he calls "the volitional dimension of commitment" and the "reasoning-centered dimension of commitment". For short, he calls these volitional commitments and reasoning-centered commitments, respectively.

Intentions, like desires, are pro-attitudes, but unlike desires, "will not merely influence my conduct, [they] will control it" (1987, p. 16). An intention not rescinded will typically be executed when the setting is right, leading to a corresponding action. This is not so with a desire. Our desires are often competing, and the object of active resistance by our will. When Bratman writes of volitional commitments, he is referring to the sense in which intentions lead to action when left in place. They *control our conduct*.

The reasoning-centered dimension of intention concerns the fact that intentions exhibit a certain *stability*. While we certainly are not *stuck* with an intention after forming one, they do "resist reconsideration" (ibid. p. 16). This stickiness is important for the role that intentions play in practical reasoning. Because intentions are sticky, we don't have to constantly keep revisiting the question of what to do. Further, their stickiness makes them good places from which we can start our reasoning, particularly when our intentions concern performing complex behaviors that requires steps of planning.

There is an obvious contrast that can be drawn between what Gilbert and Bratman say about the kind of commitment associated with intention. Gilbert appears to think that intentions are not persistent, while Bratman does.[2]

The Commitment View, which I will be sketching in the next section, will be drawing on each one of these philosophers in articulating how commitments can make constitutive principles normative.

[2] Interestingly, as we will see in chapter 4, these roles get reversed when we start talking about shared intentions.

3. Commitments and Practical Reasoning

3.1 Constructing the Commitment View

The Commitment View begins with two kinds of commitments: volitional commitments and consistency commitments. In chapter 5 of this dissertation, we will consider a third kind of commitment, what I will be calling a commitment to sociality.

Let's start with what I mean by "volitional commitments". Unlike Shpall, I will be treating volitional commitments as a kind of attitude.[3] In particular, a volitional commitment is an expression of one's will, or as Gilbert puts it, "a creature of the will". Unlike Bratman, I will not limit volitional commitments to intentions. As I will be understanding it, a volitional commitment is the building block of agency. What it is to be an agent is to be the kind of creature that makes volitional commitments. Intentions are the volitional analog of beliefs. Forming an intention, much like forming a belief, requires the capacity to form concepts and engage in first-order reasoning. I see no *a priori* reason to limit the domain of those who have volitional commitments to those with conceptual capacities or first-order reason. As such, I treat intention as a *kind* of volitional commitment. But the category of volitional commitment is much broader than intention. Any being who has genuine agential capacities forms volitional commitments. This, I want to suggest, is the hallmark of agency. The capacity to form volitional

[3] As I mentioned in section 2 above, it's actually not clear how Shpall is thinking of *volitional* commitments. He does explicitly say that commitments are normative relations, but describes volitional commitments as "constituted by your dedication" (2013, p. 727). This turn of phrase makes it sound like volitional commitments might be importantly different than the moral and rational variety in that volitional commitments are not in fact normative relations but rather are reducible to something attitudinal.

commitments is, to channel Aristotle, what distinguishes those organisms that are agents from those that are not, such as plants.

Why might we think that there can be nonconceptual volitional commitments? To answer this question let's zoom out and consider thought in general. As I will be using the term, any being who has a mind has thought. To be a subject of thought is not necessarily to be a subject of phenomenal consciousness.[4] In other words, one might think but not be *aware* that one is thinking.[5]

Following Burge (2010), I hold that thought (again, having a mind) begins at least with perception. Burge argues that objective representation of an external environment begins with perception, but perception does not require reasoning, either of the second-order or first-order variety, nor does it require conceptual capacities (which, according to Burge, are components of first-order, deductive reasoning). Perception has a non-conceptual structure; rather than conceptual, perception is demonstrative in structure.[6]

---

[4] Tyler Burge writes:

> "It is not a scientific requirement on perception that it be conscious. We know that bees and spiders have perception. We do not know whether they are conscious. Moreover, there is empirical reason to believe that some perception in bees, and in us, is unconscious. (2014, p. 401)

My claim is that there can be volitional commitments that are the analog for practical thought of perception in theoretical thought. Any being which has perceptions, in theory at least, has the mental architecture to form volitional commitments.

[5] Cf. Nagel (1974). Is there something that it is like to be a bat? Probably, but there might very well be a number of animals who have minds but lack phenomenal consciousness, and so there is nothing that it is like to be them.

[6] A lot of this is drawn from Burge (2010). For a more readable, and much shorter, articulation of the ideas in that book, I recommend Burge (2014).

Contra Burge (2010, pp. 326-341), I hold that agency requires the ability to form an objective representation of the distal environment.[7] If Burge is right that we find such representational systems only where we find perception, then this means that agency likely begins where perception begins.

Why does agency require the ability to form an objective representation of the distal environment? Answering this question is a large research project in itself. Here I can only gesture at what I think the answer is. Agency is the capacity to effect change in one's environment. But this capacity is one that is directional, not accidental. What I mean by this is that agency is *purposeful*. When one exercises agency one *aims* to bring about a change in the environment. This requires that one have a representation *as of* at least some aspect of the environment, outside of the agent itself. The point here is that agency involves intentionality about the world one is interacting with. Intentionality requires representational capacities. So agency requires representational capacities.

It should be clear how agents with sophisticated representational capacities and cognitive architecture that includes intentions are capable of satisfying the demands of genuine agency. But how about the other animals on Earth with "lesser" minds? Agency, I believe, requires two things: perception and some kind of capacity to represent *to be doneness* to the environment. If I am an organism that represents an object in my environment as food, and now I am exercising my agency to eat it, I must in some way represent that food as *to be eaten*. This is a slightly different representational capacity

[7] Burge argues that agency is pre-perceptual, that some organisms with mere sensory registration systems (such as amoeba) exhibit genuine agency. I think this is wrong for reasons that will come out in the paragraph to follow.

then perception, but there need be nothing cognitively sophisticated about it, certainly no more so than perception. And it certainly need not involve phenomenal consciousness. As far as agency is concerned, a being could have a kind of programmed script that takes a representation, e.g, *that's food,* as input, and produces a kind of executive order, e.g., <*eat that*>, as output. Such an executive order, in being demonstrative in structure and connected to perceptual representation in the way suggested, would still be appropriately representational in nature and involve the requisite connection outside of the organism that genuine agency requires. But again, none of this need have a phenomenal aspect to it.

Our subject is humans, not primitive agency. This brief discussion about the origins of agency, however, is helpful for reflecting on the nature of commitments. What I want to say about commitments should be applicable to primitive agents no less than humans. At least, that is the hope.

Let's get back to volitional commitments. A volitional commitment is, to borrow Bratman's helpful turn of phrase, a conduct-controlling attitude. It is the kind of commitment an individual makes when it exercises its agency. It is a commitment to produce a change in the world. Volitional commitments can either be present-oriented or future-oriented. Present-oriented volitional commitments involve the "here and now" execution of agency, while future-oriented commitments correspond to Bratman's future-directed planning attitudes.[8]

[8] Presumably only individuals with sophisticated cognitive capacities will possess the ability to create future-oriented volitional commitments, for this will require some degree of imagination as well as a sense of the future. It has been suggested to me by Will Swanson that jumping spiders have the ability to plan. If this is so, do jumping spiders have a conception of their future existence? I do not know the answer to this question, nor at this time do I have a conjecture as to what it might be.

The primary function of volitional commitments is the controlling of behavior in order to produce a change in the world. An important component of volitional commitments, gestured at above, concerns their representational content. Volitional commitments are *about* something. For very simple agents, the content of a volitional commitment might just be demonstrative <*eat that*> or <*attack that*>, while for a sophisticated agent the content of a volitional commitment may be quite complex, for example, <*get a degree in philosophy*> or <*deliver a paper at the APA*>.[9] What volitional commitments do is put the gears of the agent in motion toward the pursuit of the end contained in the content of the commitment. If the commitment is a "here and now" commitment, then the wheels start churning immediately. If the commitment is future oriented, then gears might be put on a time-delay, so to speak, or get one to start constructing a plan of action. Whatever the case, a volitional commitment is a commitment by an organism to exert volitional effort towards the production of some end.

The second commitment that will play an important role in the commitment-based account of practical reasoning is what I will be calling a *consistency commitment*. This commitment is similar in some ways to Shpall's rational commitment. Recall that a rational commitment is what one is committed to "having"[10] on pains of irrationality. So

[9] It will be noticed that I have not included reference to the organism in the content of the attitude. This is an intentional omission for it seems unlikely that simple agents will be able to form self-referential attitudes. Whatever the case, volitional commitments don't need reference to the individual any more than perceptions do.

[10] Shpall uses the term "having" here because the second relata in the normative relation he is concerned with is an attitude, such as a belief or intention.

if I believe that p implies q and I believe that p, then I am rationally committed to believing that q, for it would be irrational for me not to. A consistency commitment, I want to say, is similar. In fact, it has the very same implication for beliefs that Shpall's rational commitments have. I prefer the phrase "consistency commitment" to "rational commitment", for the phenomenon I want to pick out, I think, does not require second-order cognition. which is a rather sophisticated capacity that many animals lack. In fact, many animals with first-order deductive reasoning lack this capacity. Presumably, however, these beings are still *in some way* under the normative pressure to believe the consequent when the antecedent and the conditional is believed.

What I am calling a consistency commitment is a commitment to maintain consistency among one's theoretical and practical attitudes. What this entails will depend on the kind of agent in question. A being whose only theoretical cognitions are perceptions will have an easier time maintaining consistency in the theoretical realm than an individual with more sophisticated cognitive architecture that allows for the formation of complex representational states untethered to concrete reality (e.g. imagination). But a commitment to theoretical consistency is, I think, a constitutive feature of beings with practical thought. This might be surprising to those who think that the theoretical realm is cleaved off from the practical, that never the two shall meet. But non-sessile organisms are, first and foremost, practical beings. Thought develops in the service of locomotion. Thought is tethered to agency. The claim is not one about the nature of thought itself – this is not a claim about metaphysical necessity – but rather one about the nature of thought *for practical beings*. Agents have thought because of what thought enables them

to *do*, namely effect a change in their environment. Consistency among one's theoretical attitudes is conducive to survival because having veridical representations is conducive to efficacious agency, and efficacious agency is conducive to survival. The same line of reasoning applies to consistency among one's practical and theoretical attitudes. Committing to eating something that we represent as not to be eaten, committing to walking somewhere that we represent as not be entered, or approaching something that we represent as not to be approached can all lead to death.

Of course, as far as consistency is concerned, it doesn't matter which of two competing representations we give up. Sometimes, however, certain representations will not be in our power to give up. If we are unwilling to give up, say our volitional commitments, or we have two competing representational commitments that are equally recalcitrant, then our commitment to consistency will give rise to a volitional commitment to seek out adjudicatory evidence.

Now this consistency commitment is not itself an attitude, like a volitional commitment is. Rather, the consistency commitment is something more like a background framework against which thought takes place. But importantly, for my purposes, this commitment has volitional implications. When an individual falls into inconsistency, the consistency commitment moves it to do something to change the situation. This commitment to avoid inconsistency is fundamental to practical minds.

3.2 Reasoning with Commitments

This is not a dissertation about primitive agency. It is a dissertation about the foundations of human morality. An implication of this dissertation's main argument is that part of morality's foundation lies in the nature of our practical reasoning. My claim in this chapter is that an important part of practical reasoning is structured by commitments. In the rest of this chapter I'd like to make good on this claim. As such, I shall I leave primitive agency behind for the time being.

I've identified two kinds of commitments that practical thinkers have: volitional commitments and consistency commitments. On the Commitment View, practical reasoning is the process of considering the question "*shall I* ϕ?" where answering this question involves, among other things, determining whether ϕ-ing would be consistent with one's other commitments and background beliefs. When the proposed course of action is a complicated one, or when the proposed course of action reveals inconsistent beliefs or conflicting commitments, answering this question can be difficult. When this happens it will not be possible to simply determine whether the commitment is consistent with one's prior commitments and beliefs since one has just determined that there was an inconsistency in one's prior commitments and beliefs. When this happens, what one must do in order to resolve the question is, of course, seek out adjudicatory evidence.

Of course, that a certain course of action is consistent with one's prior commitments and beliefs may not necessarily settle the question of whether to ϕ, for it might be that both ϕ-ing and not ϕ-ing are consistent with one's prior commitments and beliefs. What one does here is, again, seek out adjudicatory evidence.

The best way to understand how the Commitment View works, is to see how it makes sense of the idea that constitutive principles can be normative. It turn to this issue in the next section.

4. The Normativity of Constitutive Principles[11]

In chapter 1, I defined a constitutive principle as a principle that is both descriptive of and normative for certain activities. That there could be such a principle is perplexing to some people. The descriptive and the normative reside in two different worlds. Does the constitutivist, in trying to turn descriptive principles into normative ones, turn an 'is' into an 'ought'? Does constitutivism rest on a fallacy?

Naturally, I think the answer to this question is *no*. In this section I will try to show how the Commitment View can makes sense of the normative aspect of constitutive principles. I will do so by responding to a recent objection to Korsgaard's constitutivism by Bratu and Dittmeyer (2016),[12] they wish to argue that Korsgaard—and constitutivists in general—fail to establish the truth of C2:

> **(C2)** If A is acting she should observe P, *because* observing principle P is constitutive of action.

In C2, principle P stands for the moral law. In chapter 3 I will argue that for Korsgaard, the moral law not in fact a constitutive principle of action. So strictly speaking (C2) is false.[13] But insofar as the moral law is grounded in principles that *are* constitutive of

[11] Much of this section is drawn from Bachman (2018).

[12] I will refer to them as B&D in the text.

[13] I make this argument in Bachman (2018).

action – chiefly the categorical and hypothetical imperatives – an analogue of C2 that replaces 'P' with one of these principles would be one I accept as closer to a central claim of Korsgaard's.[14] In what follows I will understand C2 as a claim about these constitutive standards. However, B&D's criticism of C2 presents a challenge to even this more modest way of understanding the claim.

Bratu and Dittmeyer suggest that C2 is an instance of what they call the *abstract constitutive principle*:

> (ACP) If a person A is partaking in a practice Y for which the observance of principle Q is constitutive, A should observe Q, because observing Q is constitutive of Y-ing. (p. 1136)[15]

B&D reject (ACP) because they think that it faces serious counterexamples, such as the following:

> *Prison Guard*: Claudia…works as a prison guard in a dictatorship. This dictatorship aggressively chases its adversaries and tries to crush their resistance by imprisoning and brutally torturing them. This is why the dictatorship expects a fair amount of cruelty of its prison guards. Claudia is told to take care of the old woman who has been leading the resistance. Having been imprisoned for some time, this woman is so frail that she will probably not survive another interrogation. Claudia wonders what she should do and after some deliberation she comes to the following conclusion: 'Claudia, you should torture this woman because you are a prison guard and where you come from a prison guard should be cruel. (pp. 1138-1139)

---

[14] A more accurate way of articulating the central claim in question would be as follows:
(C2*) If A is committed to acting, and P is a constitutive principle of action, then A must conform to P if A is to perform an action at all.

[15] ACP's analogue of C2* would be as follows:
(ACP*) If observance of principle Q is constitutive of practice Y, then in order for S to engage in Y, S must comply with Q.

B&D represent Claudia's reasoning as so:

> (1) It is constitutive of being a prison guard to be cruel to the inmates.
> (2) I am a prison guard.
> (3) If a person A is partaking in a practice Y for which the observance of principle Q is constitutive, A should observe Q. (ACP)
> (C) I have a reason to be cruel to this inmate. (p. 1139)

While B&D do not want to outright reject Claudia's conclusion—they suggest that Claudia does indeed have a *pro tanto* reason to be cruel—they suggest that any reason she *might* have stems from threats made by the regime "to attack her or her loved ones if she does not fulfill her duties as a prison guard" (ibid). The fact that torturing inmates is constitutive of being a prison guard provides absolutely no reason to torture the inmates: "considerations such as these do not carry any normative weight" (ibid).

There is a glaring problem with the example B&D chose: being cruel to inmates is in no way constitutive of being a prison guard. If it were, it would make no sense for us to praise kind prison guards as being good *qua* prison guards. But we do. And it makes sense for us to do so. So, the first premise in Claudia's reasoning is false. But this is really only a minor problem, one that can be fixed by switching Claudia's profession to *torturer*.

Does Claudia the torturer have a reason to be cruel to the old woman simply *because* Claudia is a torturer, as (ACP) suggests? B&D would say that intuitively, contra constitutivism, the answer is no:

> The fact that observing some principle is constitutive of a practice does not, in itself, speak in favor of anything, as can be shown by counter-examples such as Claudia's. Whenever it appears to be otherwise, there are always additional reasons in the offing—reasons for taking part in the practice in question—that do the real argumentative work and for which (ACP) acts as a mere placeholder. (p. 1140)

B&D suggest that Claudia's being a torturer does not, in itself, generate a reason for her to torture her prisoner; it generates a reason to torture her prisoner only if she has a *reason* to be a torturer.

There are two ways in which B&D go wrong. First, B&D fail to recognize that ACP is ambiguous; there are two ways that we can read the conditional. On one reading – call this the *wide-scope reading* – the scope of the 'should' is the conditional as a whole. On a second reading – call this the *narrow-scope reading* – the scope of the 'should' is just the *consequent*. B&D appear to embrace the narrow-scope reading of ACP. I will argue that this is the wrong way to approach ACP.

To help see the importance of this distinction, consider one of the central questions in the literature on the normativity of rationality: what is the scope of the "ought" in rational requirements such as the instrumental principle, which asks us to take the means to our end?[16] There are two ways that we can understand the scope of the "ought" in this principle:

> **IP-Narrow**: You ought to intend that M, if you intend that E and you believe that in order to E you must M.
>
> **IP-Wide**: You ought to (intend that M if you intend that E and you believe that in order to E you must M).

In IP-Narrow, the "ought" has scope only over the consequent. In IP-Wide, the "ought" has scope over the *entire conditional*. IP-Narrow thus allows us to "detach" oughts from the consequents of conditionals: when the antecedent of the conditional is true, we seem

[16] There is now a fairly voluminous literature on the normativity of rationality. See Broome (1999), (2003a), (2003b), (2005), (2007), (2013), Brunero (2008), Dancy (2000), Kolodny (2005), Lord (2014), Raz (2005a), (2005b), Schroeder (2004), Shpall (2013), Way (2010a), (2010b), (2012).

to be required, full stop, to comply with the consequent. Since the scope of the "ought" in IP-Wide is the whole conditional, we cannot "detach" the consequent when the antecedent is true. This means that there are three ways of satisfying IP-Wide: one can give up the end, give up believing the means, or intend the means.[17]

An example might help illustrate. Lisa intends to make me scream in pain. She believes (correctly) that in order to do so, she must punch me in the face. According to IP-Narrow, it would seem to follow that she ought to punch me in the face. According to IP-Wide, all that follows from Lisa's intention and belief is that she ought to either punch me in the face, not intend to make me scream, or not believe that in order to do so she needs to punch me in the face.

The difference that this example brings out between the wide and narrow readings of IP leads some to think that the wide-scope reading of rational requirements is preferable, for it allows us to avoid what some think is a pernicious bootstrapping problem.[18]

We should respond to B&D by pointing out that, much as IP is ambiguous between a narrow and wide-scope reading, ACP is similarly ambiguous between a narrow and wide-scope reading. If we accept the wide-scope reading, we can see that one

[17] Philosophers who defend wide-scope interpretations of rational requirements include Broome (op cit.), Brunero (2008), Dancy (2000), Greenspan (1975), Hill (1973), Shpall (2013), Way (2010b).

[18] For example, Bratman (1987, pp. 24-27) (in a slightly different context), Broome (2003a), (2003b), Shpall (2013). The bootstrapping worry sometimes goes by the name "the detaching problem"; see Way (2010a). Not everyone finds bootstrapping pernicious, particularly so-called 'Humeans' about practical reasons, such as Schroeder (2004). Kolodny (2005) doesn't think bootstrapping follows directly from a narrow-scope reading of rational requirements, for he thinks that rational requirements are not genuinely normative. The narrow-scope approach has also been defended by Lord (2011), (2014).

can accept ACP while also agreeing that a principle's simply being constitutive of a practice does not give one a reason to conform to it. Just as we can comply with IP-wide by abandoning our end, we can comply with ACP by abandoning our commitment to the practice. (In the case at hand, this would involve abandoning the practice of being a torturer, which is surely what Claudia ought to do.)

Let's take a closer look at the wide-scope and narrow scope readings of ACP:

> **ACP-Narrow**: Because Q is constitutive of Y-ing, person A should observe Q, if she is partaking in practice Y for which the observance of principle Q is constitutive.

Notice that with ACP-Narrow, if the antecedent is satisfied – person A is partaking in practice Y for which the observance of principle Q is constitutive – the consequent can be detached: *A should observe principle Q*. Compare this result with ACP-wide:

> **ACP-Wide**: Because Q is constitutive of Y-ing, person A should (observe Q if she is partaking in a practice Y for which the observance of principle Q is constitutive.)

With ACP-Wide, since the whole conditional is within the scope of the "should", no such detachment occurs. All that follows from the satisfaction of the antecedent is that person A should either not partake in practice Y or observe principle Q.

What we should point out to Claudia, then, is the same sort of thing we point out to any person whose end gives rise to evil means: abandon your end! While it *would* be irrational for Claudia to be committed to being a torturer *while* resisting being cruel to her prisoner,[19] this doesn't mean that what she *should* do to avoid this irrationality is to be

[19] It would be irrational because it would make her commitments incoherent. Claudia would be committed to engaging in a practice but also committed to refraining from complying with a principle she would need to comply with if she is to engage in that practice at all. This is perhaps a species of means-end incoherence.

cruel to her prisoner, or that even that it gives her *a reason* to do so. From the point of view of ACP-Wide, what Claudia *should* do is *either* give up being a torturer *or* be cruel to the prisoner. ACP-Wide itself is neutral on which she should do. But *all things considered* – that is, taking into consideration the totality of Claudia's commitments – presumably she should give up her end.

With this in mind, consider B&D's "bootstrapping" argument:

> If we accept ACP, it is possible for the dictatorship to create a reason for its prison guards to torture the inmates simply by establishing a practice which supports this kind of behavior. This shows that ACP serves as a *reasons-generating principle*, which makes it possible for us to pull ourselves up by our own bootstraps out of any normative problem: Whenever we are in need of a reason for an action x all we have to do is to point to some practice Y for which the performance of x is constitutive and then make sure that we are part of Y (1141).

As we've seen, whether one thinks that ACP generates pernicious bootstrapping depends on how one interprets the 'should' in the principle. Does the 'should' have a wide scope or narrow scope? If the 'should' has a narrow scope, ACP may very well generate pernicious bootstrapping. If, however, the 'should' has a wide scope, ACP would not generate pernicious bootstrapping.

One might think that this distinction does nothing to solve our problem, for even if we accept ACP-Wide as preferable to ACP-Narrow, it might seem that accepting ACP commits us to saying that, insofar as torturing the prisoners *is* one of the things she can do to be in conformity with the constitutive principle, Claudia does indeed have a reason

to be cruel to the inmates, simply in virtue of the fact that it is constitutive of a practice to which she is committed. But this is exactly what B&D object to.[20]

Given ACP-Wide, what Claudia ought to do is *either* torture the prisoner or abandon the practice; *the 'ought' does not distribute across the disjunction*.[21] But one might think, channeling Raz (2005b), that "people have reason to do what will bring them into conformity with reasons which apply to them" (p. 3), and so, though in general we cannot distribute the 'ought' across the disjuncts, the two disjuncts in this case are not unrelated to the source of the obligation, namely the requirement associated with ACP. Insofar as satisfying either disjunct enables us to comply with ACP, it would seem that we have a *reason* to comply with each of the disjuncts.

The problem with this suggestion is that it misunderstands the role that constitutive principles of ordinary everyday activities play in practical reasoning. The importance of establishing that a practice is governed by a constitutive principle is that it establishes what we must do *in order to engage* in said practice. But identifying a constitutive principle of a practice can just as much identify a reason *not* to engage in a practice as it can identify a reason to *do* the thing constitutive of that practice. So to identify a constitutive principle of an activity to which I am committed is to do nothing

[20] Thanks to an anonymous reviewer from *Philosophia* for pushing me to address this objection.

[21] Consider the following state of affairs, which I presumably ought to bring about: *that my children be fed a nutritious dinner*. If it is true that I ought to bring about this state of affairs, then it is also true that I ought to (feed my children a nutritious dinner or gamble away my life savings in Las Vegas). [The rule here is $Op \rightarrow O(p \vee q)$, where 'O' is the ought operator. If p is true in all "ought" accessible worlds, then (p or q) is true in all "ought" accessible worlds.]. We obviously cannot now distribute the ought across this disjunction to reach the conclusion that I ought (or even have a reason) to gamble away my life savings. [In other words, the following is not a valid inference: $O(p \vee q) \rightarrow Oq$.] *This* would be pernicious bootstrapping if *anything* is.

more than to identify what I must do to follow through on my commitment. But to do this is not to identify any reasons that I have, unless, that is, intentions give rise to reasons. But there is good reason to resist this conclusion independently of a defense of constitutivism.[22]

The point is that it is wrong to think of constitutive principles as pointing in only one normative direction, so to speak, namely toward successful engagement in a practice. According to Korsgaard's constitutivism, the normative direction in which we are to walk– that is, what we have most reason to do – is what we would endorse upon reflection; and what we would endorse upon reflection is a function of the *totality* of our commitments or practical identities.[23] Identifying an activity's constitutive principles will sometimes help us determine whether we ought to commit to engaging in said practice or not. Thus, the fact that a certain principle is constitutive of a certain activity may actually point us away from engaging in that activity, if, for example, we find out that being cruel is constitutive of that practice.

This discussion helps bring out the role of constitutive principles in practical reasoning. Identifying and conforming to the constitutive principles of ordinary everyday activities to which we are committed is a form of instrumental reasoning, since complying with such principles is a necessary means to successful engagement in the activities they are constitutive of. While such principles are normative in the sense that they designate *standards of success* for engagement in the practices they are constitutive

[22] See for instance Brunero (2009).

[23] A practical identity is a "description under which we value ourselves" (1996a, p. 101).

of, they are not normative in the sense that they *obligate us* (even *pro tanto*) without further normative support.[24] This further normative support comes from an agent's further commitments. In the case of Korsgaard's constitutivism, this story involves constitutive commitments *beyond* that of efficacy and self-constitution, most famously a commitment to valuing our own humanity (1996a, pp. 120-125), as well as contingent commitments that structure our practical identities (1996a, pp. 100-107, 128, 255-256). It is the totality of such commitments that determine what we have reason to do. Practical reasoning, so understood, is the attempt to bring newly acquired commitments into coherence with this totality.

Compare this model of practical reasoning with the model B&D propose as being "constitutive of our practice of arguing for and against something" (p. 1142):

> (NB) in order to count as valid, a normative consideration C_1 has to be supported by other normative considerations C_2-C_n. (p. 1141)

According to B&D, it is NB which explains why bootstrapping is to be avoided: "in order to count as a reason a normative consideration has to be supported by other normative considerations" (ibid.) ACP runs afoul of NB, they argue, because it "claims that constitutive standards are normative and then refuses to support this claim by pointing to further normative considerations" (pp. 1141-1142). A normative consideration, on their view, is a normative consideration only if it is supported by further normative

[24] In this respect, B&D are close to being right when they say, "The fact that observing some principle is constitutive of a practice does not, in itself, speak in favor of anything, as can be shown by counter-examples such as Claudia's. Whenever it appears to be otherwise, there are always additional reasons in the offing – reasons for taking part in the practice in question–that do the real argumentative work and for which ACP acts as a mere placeholder" (p. 1140). However, as I've tried to show, this claim is not inconsistent with Korsgaard's view, and as I will shortly suggest, there is an important difference between principles constitutive of ordinary activities and principles constitutive of agency.

considerations. While B&D fail to mention it, NB is clearly a form of coherentism, unless, that is, there can be an infinite string of normative support. Of course, what I have been defending on behalf of constitutivism in the last few pages is *also* a form of coherentism, however one that is anchored in place by *fundamental commitments*, thus avoiding, as McDowell might say, "frictionless spinning in a void".[25] Insofar as these fundamental commitments are not themselves supported by normative considerations, constitutivism so understood *does* violate NB, though not in the way that B&D suggest, for as I've been arguing, constitutive principles of ordinary, everyday activities are normative (in the sense of obligating in one normative direction) only when they are supported by the totality of one's commitments.

In this section my discussion has focused on constitutive principles of ordinary everyday activities. But for the constitutivist there is something importantly different about the principles constitutive of agency; unlike ordinary, everyday activities, the activity of agency is *inescapable* (SN p, 1). Because agency is inescapable, we cannot abandon the commitments constitutive of it in the way we can abandon our commitments to ordinary, everyday activities. For this reason, constitutivists hold that the principles constitutive of agency are "*unconditionally binding*" (SN, p. 32, emphasis added). In other words, inescapable constitutive principles don't require independent normative support.

[25] McDowell (1994/1996, p. 11).

There has been quite a bit written on whether the appeal to inescapability works in generating unconditionally binding normative principles.[26] I won't wade into this debate here.

5. The Normativity of The Instrumental Principle

In the previous section we saw that constitutive principles are a species of instrumental principles and thus that the normativity of constitutive principles reduces to the normativity of instrumental principles. But where does the normativity of the instrumental principle come from? In this section I will give a commitment-based explanation of the normativity of the instrumental principle. I will argue that the instrumental principle is a constitutive principle of action – it describes successful participation in the activity of agency – and the normativity of the principle comes from our commitment to engaging in the activity of agency. Whenever we act, we commit to complying with the instrumental principle because the instrumental principle is a constitutive principle of the activity of agency.[27]

[26] See Arruda (2017), Enoch (2006, 2011), Ferrero (2009), Silverstein (2015), Tiffany (2012), Velleman (2009, ch. 5).

[27] By "instrumental principle" I mean what in the previous section I called "IP-Wide":

> **IP-Wide**: You ought to (intend that M if you intend that E and you believe that in order to E you must M).

5.1 The Instrumental Structure of Action

Action has an instrumental structure. This is the claim central to my argument that the instrumental principle is a constitutive principle of action. It is the claim that I will be defending in this subsection.

Let's begin by fixing terminology. Following Markos Valaris (2015) I distinguish between atomic and non-atomic actions.[28] An **atomic action** is an action that doesn't require that you do something else in order to do it. A **non-atomic action** is an action that does require doing something else in order to do it. It is fairly common to call atomic actions "basic" actions. As I will use the phrase – again, following Valaris – a **basic action** is an action that one performs without having to plan how to do it. Basic actions are actions that we might have in our "repertoire" of actions, which we can execute on a moment's notice. **Non-basic actions**, by contrast, require planning on our part in order to be performed.[29] Our final piece of terminology, a **subsidiary action**, is an action that one performs as a constituent of some larger action.

Some examples will help to make these distinctions cogent. Consider the action of riding one's bicycle. This is a non-atomic action because riding a bicycle requires performing many subsidiary actions, such as pedaling, balancing, steering, and so on. What counts as an atomic action will depend on one's theory of action. For many, atomic actions are simple bodily movements, such as raising an arm, lifting a finger, tapping a desk, whistling, or nodding one's head. Some action theorists, however, have argued that

[28] Valaris, in turn, is borrowing his terminology and distinctions from Enç (2006).

[29] Again, I am following Valaris and Enç in their terminological choices here.

bodily actions are always non-atomic and that the only atomic actions are volitions, or acts of the will.[30] I take no stand in this dissertation which school of thought on this issue is correct.

Most atomic actions are presumably basic, as they typically don't require planning to perform.[31] But not all basic actions are atomic. Some rather complex actions, having been internalized, might be available to us without planning. Every morning I jump (not literally) on my bicycle and start riding to campus. Riding my bicycle is a non-atomic action, but it is something I can do without any planning. I don't have to reflect and go over what I have to do to succeed at riding my bicycle. Riding my bicycle is something that comes as naturally and automatically to me as raising my arm.

Now that we've fixed terminology, I can now explain how action has an instrumental structure. All actions are either atomic or non-atomic. Let's first consider the non-atomic case. As we've defined the terms, a non-atomic action ϕ is composed of subsidiary actions ψ_1, ... ψ_n. These subsidiary actions are *means* of ϕ-ing, for they are *how* one goes about ϕ-ing. The complex action of riding my bike is composed of several subsidiary actions (pushing the pedals, steering, braking, etc.) since they are how one goes about riding a bicycle. For any non-atomic action ϕ there will be numerous *sets* of subsidiary actions Δ_1, ... Δ_n, any of which would be sufficient for performing ϕ. I call

[30] See, for example, Hugh McCann, "Is Raising One's Arm a Basic Action" and "Volition and Basic Action" both in his (1998). McCann uses the phrase "basic action" to mean what I mean by "atomic action".

[31] This might not be true in every case. A person recovering from a stroke might have to engage in a rather sophisticated form of planning to "remember" even the most atomic of actions. McCann (1998) takes cases like this as evidence that bodily movements are not basic actions.

these sets of subsidiary actions *action schemas*. Performing ϕ requires performing one of $\Delta_1 ... \Delta_n$. Failure to perform one of $\Delta_1, ... \Delta_n$ will result in failure perform ϕ.

Now each action schema *itself* will either consist in non-atomic or atomic actions (or some combination thereof). The non-atomic actions in these schemas will themselves require performing subsidiary actions. This goes on until we reach atomic actions. The action schemas for atomic actions are the atomic actions themselves.[32]

The instrumental structure of action, then, consists in the fact that performing an action requires performing action schemas sufficient to preform the action in question.

We can now give a more precise articulation of the instrumental principle:

> **Instrumental Principle (refined):** In order for S to perform ϕ for which $\{\Delta_1..., \Delta_n\}$ is the complete set of action schemas that are sufficient for performing ϕ, then S must perform one of $\Delta_1 ... \Delta_n$.

The instrumental principle of action articulates what it means to take the means to one's end. It also establishes a standard of success for action: perform one of an action's action schema. The instrumental principle, thus, is a constitutive principle of action: it describes successful engagement in the performance of an action.

5.2 Violating the Instrumental Principle

I've just suggested that the instrumental principle is a constitutive principle of action. In chapter 1 I defined a constitutive principle of action as a principle which describes successful engagement in an activity. Failure to comply with a constitutive principle

[32] There might be an infinite number of action schemas for raising one's hand. But each action schema itself consists in nothing more than raising one's hand (in a certain way).

entails failure to engage in the activity the principle is constitutive of. If the instrumental principle is a constitutive principle of action, this would suggest that failure to comply with it will entail failure to perform an action. A lot of philosophers will initially find this to be *certainly* wrong. In her essay "The Normativity of Instrumental Reason" Korsgaard somewhat famously suggests that "we can be subject to normative principles only if we can resist them" (2008a, p. 52, fn. 39). If the instrumental principle is constitutive of action in the way I have described it, then it would appear that it is not a principle that we can resist, and if not a principle that we can resist, then not a normative principle.

Before I respond to this challenge, let me be clear as to why the instrumental principle – if it is constitutive of action in the way described – cannot be resisted. Resistance, by its very nature, is intentional. To resist something is to perform an action in defiance of it. So, to resist a principle is to perform an action in defiance of the principle, i.e., to perform an action that doesn't comply with the principle. But if the instrumental principle is constitutive of action in the way described, then by performing an action one has ipso facto complied with the principle. So one cannot resist the instrumental principle!

Can one really not resist the instrumental principle? Consider the following two kinds of cases:

(1) *Failing to perform an action at all*: One intends to ϕ, but the movements of one's body are not in conformity with the action schema one has adopted.

(2) *Failing to perform an action one intends to perform*, *but still performing an action*: One intends to ϕ but adopts an action schema that one knows to be insufficient for performing ϕ. Since an action schema itself consists in actions which themselves are composed of subsidiary actions, by discharging an

> action schema which one *intended* to discharge, one *ipso facto* performs an action and complies with the instrumental principle.

Let's consider each of these kinds of cases in turn. Case (1) is a failure to perform an action. Imagine that I intend to pick up the mug in front of me and drink from it. In order to perform this action, I must raise my arm toward the cup, grip the handle, and raise it toward my mouth. But imagine that instead of this action schema being performed, I instead begin pounding my fist on the desk. Here my body is doing something which I have not intended it to do. This is a case of bodily movement but not action. It is not action because I am not executing a schema to an intended end. If you ask me why I was banging on the table, I would likely answer, "I don't know… my arm seems to not be in my control."

Let's now consider the second kind of case. Imagine again that I intend to drink from my mug, but this time I adopt an action schema which involves me picking up the mug and pouring it on my head. I know that if I do this I will not do what I intend to do. But I do it anyways. This, one might think, is a case of resisting the instrumental principle. I am performing an action, but I am refusing to take the means to my end!

This way of understanding the second kind of case is confused. In dumping the coffee on my head, I am not resisting the instrumental principle. What I am resisting, if I am resisting anything, is drinking from the cup. Do I *really* intend to drink from the mug? Perhaps, perhaps not. The case is so strange that it is a bit unclear. But what is clear is that, whatever I intend to do, I do intend to dump the contents of the mug on my head. And in doing this, I am complying with the instrumental principle for I am executing an

action schema in the pursuit of an end.[33] In this case it is true that I am acting, but false that I am resisting the instrumental principle. In fact, I am acting precisely in virtue of the fact that I comply with the instrumental principle. I take the means to my end.

Now, there are what we might consider intermediary steps between these two cases. I might, for instance, intend to drink from the cup, but fail to get the contents of the cup to my lips and instead pour it all over the front of my shirt. How does my account of the instrumental principle handle this case?

We might be tempted to say that this is a case of action where one does not comply with the instrumental principle. It is action because I have performed at least some of the subsidiary actions in the action schema I've executed, but not a case of complying with the instrumental principle because I have failed to take the means to my end.

Here, again, I think we've made a mistake in the interpretation of the case. Pouring the coffee down my shirt is certainly not an action that I've intended to perform. What I have done in this instance is failed to fully execute an action schema that I have adopted. So am I acting? Well it depends upon the scope of the action. At certain points in the chain of the behavior it would be correct to say that I am performing an action. When I raise my arm toward the cup I am performing an action. When I grip the mug I am performing an action. When I raise the mug toward my mouth I am performing an action. And it is important to note – before we move further down the chain – that at each

[33] The end is pouring the contend of the cup on my head. The action schema executed involves the subsidiary actions of raising my hand toward the cup, lifting it to above my head, tipping it, etc.

one of these steps I've complied with the instrumental principle. But how do we describe the behavior of pouring the coffee down my shirt? At a certain level of description this is an action, but at another level it is not. If we describe what I am doing as *pouring coffee down my shirt*, then it would be false to ascribe this to me as something that I intend to do. But if we describe what I am doing as *tilting the mug so that the contents can come out* then this *is* an action that I am performing. And *again*, in doing this I am complying with the instrumental principle, for I am performing subsidiary actions that constitute an action schema I am executing.

My interlocutor might insist: "But haven't you failed to take the means to your end? After all, you didn't drink from the mug, which was your end. Instead, you poured it all over your shirt like a damn fool. And since you failed to take the means to your end, you failed to comply with the instrumental."

My response: In the actions that I performed (raising my arm, gripping the handle, lifting the mug, etc.), in performing them I complied with the instrumental principle. In the action I failed to perform (drinking the coffee), I failed to comply with the instrumental principle.

The claim that I am defending is that the instrumental principle is a constitutive principle of action. This means that in order to perform an action we must comply with the principle. Failure to comply results in failure to perform an action. In the case where I pour coffee down my shirt, there are actions that I perform. In performing them I comply with the instrumental principle.

All this being said, I think there is *something* to Korsgaard's claim that a principle is normative only if it is possible to resist it. The truth to this claim lies not in the idea of resistance, but rather the underlying idea that normative principles must be *violable*. That is, a principle is normative only if it is possible to fail to comply with it.

As we have seen, it *is* possible to fail to comply with the instrumental principle. Of course, this means that one has failed to act. Presumably it is possible to do this.

5.3 The Normativity of the Instrumental Principle

I have argued that the instrumental principle is a constitutive principle of action. It describes how to successfully engage in the activity of action. I have promised that the commitment account of practical reasoning can explain the instrumental principle's normativity.

In section 3, I suggested that constitutive principles become normative for an individual when one commits to engaging in the activities the principles are constitutive of. The instrumental principle is a constitutive principle of action. So the principle becomes normative for an individual when one commits to engaging in the activity of action. An individual commits to engaging in the activity of action when if one intends to perform an action. The explanation for the normativity of the instrumental principle, then, is that it binds to an individual in virtue of that individual committing to perform an action.

6. Conclusion

In this chapter I laid out the elements of the Commitment View of practical reasoning. According to this view, a core component of practical reasoning concerns fitting our commitments and beliefs into a coherent web. Now, in the scope of this dissertation, the Commitment View is really just a means to an end. Our aim is to articulate and defend a certain conception of constitutivism – what I call *Social Constitutivism*. How does the Commitment View help us get there?

First, the Commitment View tells us how constitutive principles become normative. Constitutive principles, we have seen, are a species of instrumental principles. They become normative for an individual when the individual commits to a practice of which the principle is constitutive.

Second, the Commitment View provides us with a framework for working out a viable version of constitutivism. If constitutive principles describe successful engagement in certain activities or practices, and being committed to engaging in a practice governed by such a principle is what makes constitutive principles normative for us, then if the moral law is to be a constitutive principle, the Commitment View tells us that we must be committed to some practice of which the moral law is constitutive.

Third, the Commitment View tells us that agency is governed by constitutive principles. Action, we saw, is governed by the Instrumental Principle. The Commitment View tells us that it's normativity can be explained by our commitment to engaging in the activity of agency. Might there be a form of agency of which the moral law is a constitutive principle?

Fourth, the Commitment View tells us how morality might be universal and categorical. If the moral law is a constitutive principle of an activity to which we are inescapably committed, then it will be normative for everyone regardless of what contingent inclinations they might happen to have.

The Commitment View, I want to suggest, is a potentially powerful tool in the hands of a constitutivist.

Chapter 3: The Bad Action Problem and the Structure of Constitutivism

1. Introduction

> Recall our definition of constitutive principle from chapter 1:
>
> A **constitutive principle** is principle that is descriptive of and normative for engagement in an activity or practice.

In chapter 2, I argued that the instrumental principle was a constitutive principle of action. We saw that one implication of this claim was that every action conforms in some way to the instrumental principle. In this chapter we turn to an objection to constitutivism related to this idea.

Paul Katsafanas (2013) and Matthew Silverstein (2016) have argued that some versions of constitutivism, in particular that advanced by Christine Korsgaard (2009), lead to a counterintuitive result: it is impossible to *act immorally*. Consider the game of chess.[1] The reason that the rules of chess bind chess players is that conforming to the rules of chess is *necessary* in order for one to count as *playing chess*; if one's actions fail to conform to chess' rules, then one's actions are not *chess* moves. But if morality is to agency as the rules of chess are to chess, then this would seem to suggest that one's movements count as actions only if one conforms to the rules of morality. And *this* would seem to imply that one simply cannot *act* immorally. Call this *the bad action problem*. In general, a version of constitutivism suffers from the bad action problem if the standards it

[1] The objection I'm about to raise is not the one advanced by either Katsafanas or Silverstein. Rather, it is meant as more of a prima facie problem – an intuitive way to see the issue – that doesn't require understanding the complicated details that will be the focus of this chapter.

posits for good action are the same as – or can be reduced to – the standards that distinguish action from mere behavior. In such a theory every action will be good and there can be no bad actions.

My aim in this chapter is to show how Korsgaard can solve her bad action problem. In the course of doing so we will discover two things. First, that many criticisms of Korsgaard proceed from misunderstanding the structure of her view. Clearing up this misunderstanding clears up most charges that she succumbs to the bad action problem. Even so – and this is the second discovery – a residual bad action problem remains for Korsgaard. I will argue that ultimately her bad action problem results from her theory of functions. By tweaking this theory in a sensible way, her bad action problem will completely dissolve.

But why study the bad action problem? As will become clear in the course of this chapter, solving the problem requires getting clear about the underlying structure of constitutivism. So while much of this chapter will be focused on the details of Korsgaard's view in particular – getting clear on these details is worth doing for it's own sake – I will suggest at the end how we can generalize the results for thinking about constitutivism in general, and identify a path forward for those who find the structure of Korsgaard's view appealing, but not the metaphysics of self-constitution. This, I will suggest, is the path of *Social Constitutivism*.

2. Preliminaries

Why in general is it thought that constitutivism faces the bad action problem? The issue stems from the problem that constitutivism is trying to solve: *how can there be requirements that are universally and categorically binding?* The constitutivist's solution, of course, is to ground these requirements in features constitutive of human agency. If F is a constitutive feature of human agency, then all human agents will exhibit F. If F gives rise to moral requirements, then such requirements will be binding for all human agents, since all humans exhibit F. The bad action problem is generated when we ask what the relationship is between F and the requirements of morality.

One way this story can go is for the constitutive features of agency to generate what Andrews Reath (2010) calls "formal principles":

> [F]ormal principles… are uniquely suited to apply with normative necessity to some domain of rational activity. We may think of the *form* of some rational activity or object of cognition as the constitutive or defining features of an activity or entity of that kind – the features that an activity or entity must possess to count as an instance of that kind. The form of some rational activity or object of cognition will be associated with a *formal principle* that is constitutive of that rational activity or object of cognition. The formal principle of some rational activity would be the guiding internal or constitutive norm that a **subject must follow in order to engage in that activity.** By specifying the form of that activity, it provides a norm that **anyone** engaged in that activity **must satisfy** and that in some sense **does guide** any instance of the activity **(even if defective).** (p. 42, bold emphasis added)

Chess, we have seen, is defined or constituted by a set of rules. What counts as a move *in the game of chess* is defined by these rules. If I move my knight up three and over five then what I have done is not a *chess* move. To use Reath's language, the rules of chess are chess' formal principles, and I must follow them in order to engage in the activity of chess.

As Reath is quick to note, constitutivism, so understood, faces a problem:

> "If a formal principle is constitutive of an activity, it must be implicated in all instances of the activity. **The problem is how to characterize mistakes or defective instances of an activity.**" (ibid, fn. 20; emphasis added)

In order to make a move in chess, I must conform to the rules of chess; non-conformities are not chess moves. If the principles of morality are to agency what the rules of chess are to the game of chess, then it would appear that I must conform to morality in order to exercise my agency. In other words, this idea leads us straight into the bad action problem.

The problem Reath notes here is not a new one for constitutivism. Railton (1997) notes a similar problem with the constitutivist turn:

> Eager for a secure justification … we sought a requirement – a 'must' – that applies non-hypothetically, arising from the very conditions of agency. That now looks unwise. For then there could be no such thing as failure to conform *on the part of an agent*. (p. 71)

The bad action problem is still thought to be a relatively live one for constitutivists. In a recent paper, Matthew Silverstein (2016) argues that

> if the authority of the standards governing action springs from what is constitutive of agency, then any agent will necessarily act in accordance with those standards merely by virtue of being an agent. There seems to be no room for the possibility of bad or incorrect action. (p. 217)

One might think that given the bad action problem, constitutivism is a non-starter of a view. I think this is too quick. But I am not the only one. Some theorists friendly to constitutivism have suggested that the solution to the bad action problem lies in

recognizing that the constitutive features we want are not formal principles, but rather constitutive *aims*.[2]

As I explained in chapter 1, a constitutive aim of an activity is an aim that an individual has to adopt (in some sense) in order to count as engaging in that activity.[3] Consider archery. In order to engage in archery an individual must aim to hit the bullseye. If one doesn't adopt this aim, one is merely shooting an arrow. So hitting the bullseye is archery's constitutive aim. But notice that one can *aim* to hit the bullseye without succeeding in so doing: one can miss. But insofar as one is *trying* to hit the bullseye, one is engaging in archery; one need not actually *hit* the bullseye in order to do so. Of course, hitting the bullseye is a standard of success, so if you do so, your shot is good.

Many think that chess' constitutive aim is checkmating your opponent.[4] But is this so? I can play chess with many different aims. I can, for example, play with the aim of teaching my niece, play with the aim of passing away the time, and so on. Insofar as I'm following the rules of chess, it's not obvious that I need to aim at actually *check-mating* my opponent. Of course, check-mating one's opponent is in an important sense the "goal" of chess. But it is not clear that it is *essential* to my playing chess that I aim at

---

[2] The constitutive aim strategy is adopted by Velleman (2000, 2009), Katsafanas (2013), and, to a certain extent, Matthew Silverstein (2016), though Silverstein ultimately argues that a successful constitutivist strategy must rest on a reductive account of normativity. He makes a similar claim in Silverstein (2012). Korsgaard (2009) is sometimes described as one who adopts a constitutive aim strategy (see, for instance, Enoch 2006, p. 179), but both Katsafanas (2013, pp.45, 86-87) and Silverstein (2016), correctly in my opinion, suggest that Korsgaard is not best understood in this way.

[3] The "aiming" need not occur at the "agential" or "personal" level. See Velleman (2000, pp. 19-21).

[4] See, for example, Katsafanas (2013, pp. 1-2, 38-41, 179), Ferrero (2009, p. 306), Enoch (2006, p. 185; 2011, p. 211).

checkmating my opponent. As long as I follow the rules of chess, it would seem that I count as playing chess.

Whether chess does indeed have a constitutive aim is not a deep or interesting question. We can grant for the sake of argument that at least some activities have constitutive aims, but we should recognize that some activities don't. The interesting question is whether agency is the kind of thing that has a constitutive aim and whether this can indeed help us solve the bad action problem.

Katsafanas (2013, pp. 61-63) explicitly identifies the constitutive aim strategy as a solution to the bad action problem. Silverstein (2016, pp. 228-230) sees a similar benefit to the constitutive aim strategy: "Any view with this structure will leave room for behavior that is governed by the aim (and so counts as action) but that fails to achieve that aim (and so counts as incorrect or defective)" (p. 229). Given these statements, one might think that there is nothing problematic about appealing to constitutive aims. However, as Silverstein points out in a footnote, Clark (2001) successfully presses the bad action problem against Velleman (2000, ch. 8), who argued that autonomy, understood by Velleman as conscious control of one's behavior, is the constitutive *aim* of "full blooded" intentional action.[5] On this earlier version of Velleman's view, this aim had to be satisfied in order for a behavior to count as action. For some activities – walking, skipping, catching a ball, making a mess, writing a sentence, making a wish, kissing, hugging, getting dressed, tying your shoes – the standard of success (and so constitutive aim) is also the individuating condition. These activities don't allow for

[5] In response to Clark, Velleman changes his view in the introduction to his (2000).

defective instances. When we evaluate these activities as good or bad we often apply *external* standards when evaluating them, not ones that are grounded in the nature of the activities themselves. A bad dresser is not one who has trouble getting his shirt on – a person who hasn't gotten his shirt on *has not dressed*, and so is not a *bad* dresser – but rather one who picks unfashionable outfits. The standards of fashion are external to the nature of getting one's clothes on.

This is even clearer with walking. To walk is to put one foot in front of the other in a measured pace. This is the constitutive aim of walking. But notice that it is also the standard of success for walking in the sense that if one doesn't put one foot in front of the other in a measured pace one isn't walking. When an activity has success *baked into* the nature of the activity, that activity's constitutive aim will be satisfied in every instance. In his (2011) paper, Clark argued that Velleman's (1996) description of agency as aiming at conscious control of one's behavior turned agency into one such enterprise. Ascribing a constitutive aim to agency does not, itself, provide us with a solution to the bad action problem

There is another *prima facie* problem with the constitutive aim strategy. Recall that the constitutivist wants ultimately to derive morality out of constitutive features of agency. One easy way to do this is to make the moral law the constitutive aim of agency. For obvious reasons this would be a nonstarter of a view, but play along with me for now. A view with the moral law as the constitutive aim of action would not fall prey to the bad action problem, for presumably on such a view one can aim to conform to the moral law while falling short of doing so. However, such a view would have an odd upshot:

immoral action occurs when one *tries* to act morally, but fails to do so. While it might be true that when we act immorally we see some *good* in so acting – this is the thesis that sometimes goes by the name "the guise of the good" – it seems laughable to suggest that I'm aiming to conform to the moral law when I steal your wallet.[6]

No constitutivist thinks conformity to the moral law is the constitutive *aim* of action. But once we recognize the absurdity of such a view, one might begin to wonder about the constitutive aims posited by the other views. According to Velleman (2009), the constitutive aim of agency is intelligibility or self-understanding. Do we really aim, at *self-understanding* whenever we act? According to Katsafanas (2013), the constitutive aim of agency is will-to-power. Do we really aim at *power*, as Katsafanas suggests? What, exactly, does it even *mean* for action to have an aim? This is a difficult question. While it has been discussed at length, the answers that have been proposed are not all that clear.[7] One might be tempted to draw, as Velleman has, from the literature on the nature of belief, which holds that belief aims at truth; however, it's not immediately clear how this will be beneficial to the constitutivist's cause. Consider, for example, one of the more popular accounts of what it means to say that belief aims at truth. According to this view,

[6] Setiya (2007) argues that constitutivists, or "ethical rationalists" as he calls them, are committed to some form of the guise of the good thesis. The arguments here are not meant to be an attack of this thesis, though like Setiya, I find the thesis dubious. However, I'm not convinced that the thesis is one constitutivists need be committed to.

[7] For discussion of this question see, for example, Velleman (2000, pp. 15-31), (2009, pp. 134-147); Katsafanas (2013, p. 37-41, 61-65). There has also been a robust discussion of a parallel question in the metaphysics of belief: what does it mean to say that belief aims at the truth? For canonical defenses of the thesis that belief aims at truth, see for example, Shah (2003), Shah and Velleman (2005), Velleman (2000 pp. 244-281), Wedgwood (2002), and Williams (1973/1995). For the controversy surrounding the thesis see Chen (2013).

sometimes called "normativism", to say that belief aims at truth is to say that truth is the standard of correctness for belief.[8] To see the implications of this view for our project, notice what this view *doesn't* say, namely that truth is the standard of correctness for belief *because* belief aims at truth. Rather, normativists hold that the explanation for belief's aiming at truth is the fact that truth is belief's standard of correctness. This won't help our constitutivist, who wants to *derive* action's standard of correctness from action's constitutive aim, not vice versa.

In order to employ the aiming strategy, the constitutivist must give us an independent account of what it means for action to have a constitutive aim. It's not that constitutivists haven't tried.[9] The problem with the constitutive aim strategy is rather that the stories we are told about action having a constitutive aim are not all that compelling. A central problem concerns our experience of acting. Is there really a common thing we all aim at whenever we act? In an attempt to skirt around this problem, Velleman (2000) has suggested that action's constitutive aim is a *sub-agential* motive; it is not a motive ascribable to an individual at the personal level, but rather is a motive that regulates sub-agential behavioral mechanisms. While pushing the issue down a level might answer the question of how there could be a constitutive aim of action if there seems to be no common aim to our actions, doing so doesn't solve the more pressing problem of what it

[8] Wedgwood (2002), Shah (2003).
[9] See, for example, Velleman (2000) and Katsafanas (2013).

means for our action to have a constitutive aim. If anything, making the aim sub-agential wraps the idea in obscurity.[10]

The critical comments presented in this section are not meant to be a knock-down argument against the constitutive aim strategy. My point, rather, is that turning to constitutive aims to solve the bad action problem is not obviously a winning strategy for the constitutivist. At the very least, there are costs associated with doing so. I think this is reason enough for a constitutivist to explore her other options. Whatever the case, the solution to the bad action problem I propose in what follows makes no reference to the notion of a constitutive aim.

The rest of this chapter will be devoted to discussing the bad action problem as it applies to Korsgaard's version of constitutivism. As I mentioned in the introduction, I do this for two reasons. First, the underlying structure of her view, which I will call the *commitment strategy*, has I think not been fully appreciated by commentators. Second, Korsgaard's view is one that has sustained the most detailed criticism vis-à-vis the bad action problem. I believe that we can learn a lot about how constitutivism works by investigating the underlying structure of Korsgaard's view. I will show that we can solve the bad action problem even if we adopt the constitutive principle strategy.

[10] Unfortunately, Katsafanas (2013) is no more helpful, for he takes an uncritical stance toward the idea of actions having a constitutive aim.

3. Korsgaard's Bad Action Problem

Recall that a constitutivist suffers from the bad action problem if the standards for *good action* are the same as the standards that distinguish action from mere *behavior*. Such an account would leave no conceptual space for bad actions.

Why is it thought that Korsgaard falls prey to the bad action problem? It should be noted that Korsgaard herself seems to suggest a lurking problem. When discussing the teleological foundations of agency, Korsgaard illustrates her view with a discussion of house building:

> If you fall too short of the constitutive standard [of a house], what you produce will simply not be a house. In effect this means that even the most venal and shoddy builder must try to build a good house, for the simple reason that there is no other way to try to build a house. Building a good house and building a house are not different activities: for both are activities guided by the teleological norms implicit in the idea of a house. Obviously it doesn't follow that every house is a good house, although *there is a puzzle about why not*. (2009, p. 29, emphasis added)

Presumably, the activity of building a house represents acting in general. When Korsgaard says that you will not build a house if what you build falls too short of the standards of a house, she is suggesting that if your behavior falls too short of the standards of an action your behavior will not count as an action. In considering this, let's start by looking at how the house example works and then proceed to the case of action.

Korsgaard suggests that "a thing comes to be, as the kind of thing that it is, when a certain form is imposed on matter" (2008, p. 135). She continues, "The thing is what it is when its parts are arranged in a way that makes it capable of the activities that are essential to or characteristic of it—capable of performing its function" (ibid). Korsgaard, it would seem, individuates objects according to their form. This means an object is a

house in virtue of the fact that it is formed in a way conducive to providing shelter.

But if an object is a house in virtue of having a form that enables sheltering, and a house is good when it enables sheltering, then all houses will be good houses and there will be no bad houses. So engaging in the activity of building a house requires building a good house, for if I build a bad house, I will be building something lacking the form of a house, and so will not be building a house at all.

We can apply this line of thought to the exercise of our agency. When I "build an action", I must build something with the form of action. Since the form of a thing is the organization that enables it to perform its ergon, and the form of action is action's principle, the form of action will be that principle which enables efficacy and self-constitution. According to Korsgaard, that principle is the moral law. It would appear, then, that in order to act I must adopt the moral law as my principle. If this is the case, then it's not clear how it will be possible to "build" a bad action.

Katsafanas (2013) has advanced the bad action problem in a different way. His argument comes in the form of a dilemma: either Korsgaard faces a bad action problem or she makes the moral law optional.

The first horn of the dilemma gets started by reflecting on the following passage from *Self-Constitution*:

> You must operate as a whole, as something over and above your parts…and in order to do this…you must will your maxim as universal laws. (Korsgaard, 2009, p. 72; cited in Katsafanas, 2013, p. 95)

Here is Katsafanas commenting on this passage:

> The CI [categorical imperative] tells us to do just that: act only on those maxims that we can will as universal laws. So, in the above passage, Korsgaard contends

> that if we are to operate as unified wholes, we must act on the CI. Principles that are inconsistent with the CI either fail to unify agents at all, or fail fully to unify the agent. (2013, p. 95)

Katsafanas' thought here is that insofar as action requires unification, and we are only unified when we act on the CI, we only truly act when we conform to the CI. But insofar as the CI is also the standard for good actions, only good actions will count as actions.

Katsafanas suggests a way out: maybe unification, and thus action, comes in degrees. As we've seen, principles can unify us to different extents. Principles that fail to *fully* unify us rely on the obtaining of contingent circumstances to hold us "together." But when the circumstances present themselves, I will be unified enough to act. Good actions will be those that are fully efficacious and fully unify the agent, while bad actions will be those that are only *partially* efficacious and provide only *partial* unification.

The problem with this solution is that morality is now optional:

> [I]f all that it takes to perform an action is to act on a principle that unifies the agent to some extent, then our commitment to action yields only a commitment to acting on principles that unify us to some extent, not to acting on principles that unify us to the fullest extent. (2013, p. 101)

So if we don't have to be *fully* unified in order to perform an action, then it would appear that acting *itself* doesn't commit us to acting in accordance with the moral law. This is not a good result if we wish to show that the universality and categoricity of morality can be explained by the fact that we are committed to morality simply in virtue of our commitment to agency.

We can summarize Katsafanas' dilemma as so: either action requires conformity to the CI or it does not. If it does, then there is no room for bad action; if it does not, then morality is optional, and so not categorical and universal.

4. Stop One *En Route* to a Solution: Distinguishing the CI and the Moral Law

In *SN*,[11] Korsgaard defends a somewhat idiosyncratic interpretation of Kant's Universal Law formulation of the categorical imperative. The CI "tells us to act only on a maxim which we could will to be a law" (1996a, p. 98). But what, exactly, does this mean?

> [C]onsider the content of the categorical imperative, as represented by the Formula of Universal Law. The categorical imperative merely tells us to choose a law. Its only constraint on our choice is that it has the form of a law. And nothing determines what the law must be. *All that it has to be is a law*. (ibid)

Here, Korsgaard is "[making] a distinction Kant doesn't make" (ibid). The CI, on this understanding, tells us merely to choose a *law*. The *moral* law, on the other hand, tells us to choose a law that "all rational beings could agree to act on together in a workable cooperative system" (1996a, p. 99).

Korsgaard notes that "any law is universal," as this is just what the form of a law is, but the universality of a law does not "settle the question of the *domain* over which the law of the free will must range" (ibid):

> If the law is the law of acting on the desire of the moment, then the agent will treat each desire as a reason, and her conduct will be that of a wanton. If the law ranges over the agent's whole life, then the agent will be some sort of egoist. **It is only if the law ranges over every rational being that the resulting law will be the moral law**. (ibid; emphasis added)[12]

In more recent work, Korsgaard still embraces this distinction. In *Self-Constitution*, Korsgaard writes:

---

[11] The argument is also made in Korsgaard (1996b, p. 166).

[12] It is debatable whether this is an adequate reading of Kant. Reath (2006, p. 164, n. 16), for instance, has argued that universality is much narrower for Kant than it is for Korsgaard. What matters for our purposes is not whether Korsgaard has given us an adequate interpretation of Kant, but rather getting clear on her view.

> To get from the categorical imperative to the moral law, two more things are necessary. First of all, we must establish that the domain over which the universal law ranges must be rational beings as such: that is to say, when you will your maxim as a universal law, you must will it as a law for every rational being. And second, we must establish that the reasons embodied in universal maxims must be understood as public, or shareable reasons: reasons that have normative force for all rational beings. (p. 80)

In *Sources of Normativity*, Korsgaard identified the CI as the law of a *free will*—it tells us how we must act in order to act *freely*. In *Self-Constitution*, Korsgaard makes a stronger claim: "The argument identifies the categorical imperative as a constitutive principle of volition, so it is about what we must do insofar as we act at all" (ibid). Conformity to the categorical imperative, it turns out, is constitutive of action *as such*.

Why am I making so much hay out of this distinction? First, while for Korsgaard acting as such does require conforming to the CI, it does not require conforming to the moral law. Second, though conformity to the moral law is not required for acting as such, we will see that acting as such *commits* us to complying with the moral law. This move will help us secure the non-optionality of morality while leaving room for bad actions. Finally, a central problem with Katsafanas' dilemma is that he overlooks the distinction between the CI and the moral law and so also overlooks the solution to his dilemma.

5. Stop Two *En Route* to a Solution: Individuating Functional Kinds

The second step in our solution to the bad action problem requires that we make an important adjustment to the way that Korsgaard goes about answering the following question: what makes a token object O an instance of the functional kind K?

In *Self Constitution*, when introducing her teleology, Korsgaard states that, "what makes an object the kind of object that it is—what gives it its identity—is what it does, its *ergon*." (2009, p. 27). Later, however, she says that "in order to establish what the constitutive standards of anything are, we must look to its ***form*** in the Aristotelian sense—to ***the teleological organization that makes it the kind of thing that it is***" (2009, p. 82; emphasis added in bold). The former passage suggests that we individuate according to function, while the latter passage suggests that we individuate according to form. Which is it?

Korsgaard's settled view seems to be that function (or ergon) and form are not quite two different things. In *Self Constituion,* we've seen that Korsgaard unpacks 'ergon' as "purpose, function, or characteristic activity". In her paper "Aristotle on Function and Virtue" (collected in 2008), Korsgaard digs a bit deeper. There, she suggests that "the form is the function of a thing" (2008, p. 137). Korsgaard suggests that the best way to understand function is in terms of "how a thing does what it does" (2008, p. 138), as this gives us the best way of understanding Aristotle's claim that to know a thing is to know its form. A person who understands how a thing is put together without knowing what the thing does can hardly be said to understand the thing. So a thing's form can't merely be how that thing is put together. But merely knowing what a thing does is also insufficient for understanding the thing. Rather, to truly understand a thing requires understanding not just what it does or how it operates, but understanding *how* it does what it does. Knowing a thing requires understanding how a thing's organization *enables* it to operate.

Though this view may help us understand Aristotle, I think we are not left with a good theory of functions, primarily because the view leaves us with no conceptual space for *malfunctions*. When we ascribe a function to an artifact, we are typically ascribing a *purpose* to it. The function of an artifact is that effect the thing is *supposed* to produce. But sometimes artifacts fail to produce their desired effect. For example, I once bought a toaster from Sears, which, regardless of what setting I put it on, would produce toast that was burnt to a crisp. This is not what the toaster was *supposed* to do. To be sure, though, burning toast was something that my toaster *did*. And there was a *way* that it did this. But burning toast in the way that it did was not my toaster's *function*. What it was supposed to do was *toast* my bread, not *burn* it.

Korsgaard's teleology faces a serious difficulty when it comes to accounting for malfunctioning items, such as my toaster. If things are individuated according to *how they do what they do*, then since a "malfunctioning" thing produces a different effect than a "properly functioning" thing—and for both we can say that there is a way that they do these different things—we seem to be forced into saying that the two things are of different kinds. The bread burner, in having a different form, is a different *kind of thing* from the bread toaster, and so we can't say that the former is a malfunctioning toaster at all.

Here's what we should we say about malfunctioning artifacts: an artifact malfunctions when it fails to fulfill its function, and when the explanation for the failure is rooted in the artifact's form. My toaster burned my toast because of the way it was put together. Maybe the temperature control knob was not properly constructed. Whatever

the exact explanation, it was because of the way that my toaster was put together that it could not do what it was *supposed* to do. It did not have a form conducive to function fulfillment.

Making sense of malfunction requires that we pry apart form from function. In order to say that my toaster is malfunctioning, there must be something that it is *supposed* to do. This thing is its function. So my malfunctioning toaster has the function of toasting (but not burning) bread. But my toaster does not have the proper *form* of a toaster. If it did, it would be able to toast my bread without burning it. It would seem, then, that insofar as my malfunctioning toaster *is* a toaster, what makes it a toaster must be its function not its form.

It is not an uncommon desideratum to require that a theory of functions make room for malfunctions; it's one that many philosophers of biology have. Much as I've argued with respect to artifactual functions, many philosophers of biology think that making room for malfunctioning traits requires that traits be typed by reference to their function rather than form. Consider the following passage from Karen Neander:

> 'Heart' cannot be defined except by reference to the function of hearts because no description purely in terms of morphological criteria could demarcate hearts from non-hearts. … Highly significant, moreover, is that for the purposes of classifying hearts what matters is not whether the organ in question manages to pump blood, but whether that is what it is supposed to do. The heart that cannot perform its function (because it is atrophied, clogged, congenitally malformed, or sliced in two) is still a heart. (1991: 180; quoted in Neander, 2002: 391)

Here, Neander makes a similar argument about hearts that I made above about toasters: functional kinds are properly individuated according to function, not form.

On this Korsgaardian (but not Korsgaard's) view, what makes an instance of

behavior an action is not the principle that organizes it, but rather its function or purpose. The function, as we've seen, is efficacy and self-constitution. But how does an instance of behavior "count" as an instance of the kind *action*? To put the question another way, how does an instance of behavior come to have the function of *efficacy* and *self-constitution*? Action is an interesting case because it is both a natural function and an artifactual function. It is natural, because action is an exercise of our agential capacities, a trait selected for by natural selection. But it is artifactual because action is an activity we engage in *intentionally*. But whether we view action as a natural trait or an artifact we arrive at the same thing. As a natural trait, actions get their function from the function of the capacities that they are exercises of, and the function of our agential capacities is to *do stuff*. As an artifactual function, actions get their function from the user's *intentions*, that is, how we *use* our actions, but we use our actions *to do stuff*. So on either path we take, whether we consider actions as natural traits or artifacts, an instance of behavior is an instance of the kind *action* if it is appropriately tied to our agential capacities, namely by way of our intentions, an aspect of which, as we saw in chapters 1 and 2, is a commitment to *doing stuff*, or efficacy.

6. Responding to Katsafanas' Dilemma: Horn 1

According to the first horn of Katsafanas' dilemma, performing an action requires conforming to the CI, but conforming to the CI leaves no room for bad action. Katsafanas reads Korsgaard as giving us the following argument:

a. An agent's A-ing is an action iff A-ing is attributable to the agent as a unified whole.

b. An action is attributable to the agent as a unified whole iff the principle of the agent's action renders the agent [efficacious].[13]
c. The CI is the only principle which renders the agent [efficacious].
d. Therefore, an agent's A-ing is an action iff the principle of the agent's A-ing is the CI. (Katsafanas: 97)

The problem with this argument is that (d) seems to leave no room for bad action:

> An action is good iff its principle is the CI; by (d), A-ing is an action iff its principle is the CI; therefore, every action is good. In other words, (d) entails that the constitutive principle is achieved in each instance of action. If no intentional action can fail to manifest the constitutive features of action, then no intentional action can be defective as an action. (Katsafanas: 98)

In a certain sense, Katsafanas is right that for Korsgaard, efficacious agency requires conformity to the CI. But this is not because, as Katsafanas suggests, conformity to the CI *as such* renders the agent a unified whole. As we saw in the previous section, the CI as Korsgaard understands it merely tells us to act in accordance with a principle, but it doesn't tell us *which* principle to act on. The only principle which, *as such*, renders an agent an efficacious, unified whole is *the moral law*. With this point in mind, consider the following passage from Katsafanas:

> In short, [Korsgaard's argument] is as follows: if I choose some principle other than the CI, then any [efficacy] that I seem to exhibit will be purely accidental; it could dissolve at any time. But, if this happens—if I choose a principle that potentially compromises my [efficacy]—then I am not really unified at all. (pp. 95-96)

Katsafanas cites p. 78ff of *Self-Constitution* as evidence that this is Korsgaard's view. In this portion of *Self-Constitution*, Korsgaard is in the midst of denying the possibility of *particularistic willing*, i.e., acting on principles that are not universal in *any* sense (and so

[13] I have replaced 'capable of diachronic stability' with 'efficacious.' I do this because 'rendering efficacious' is the phrase that Korsgaard uses and as far as I can tell Katsafanas is using 'diachronic stability' to mean the same thing.

not acting on a principle at all). Without a principle, we will be unable to bridge the gap created by reflection, and so even a *momentary* efficacy is impossible, for without our endorsement of a principle, there is no *you* over and above your inclinations, and so no *agent* to whom the action would be attributable. But if particularistic willing is impossible, this means that willing necessarily has a universal scope to it, and so willing necessarily involves adopting a principle.

So it would seem that (d) is true, at least under a certain reading. However this does not lead to the conclusion that there are no bad actions, for as we've seen, not all principles unify us and so enable efficacy to the same extent. It's simply not true that "an action is good iff its principle is the CI." But why would Katsafanas *think* it true? Katsafanas is aware that different principles unify agents and thus make them efficacious to different extents (Katsafanas: 91-95). When discussing this aspect of Korsgaard's view, Katsafanas cites passages in the second half of *Self-Constituion* (133ff.), where Korsgaard turns to a discussion of how the moral law better unifies an agent and contributes to greater efficacy than other principles. Katsafanas takes her discussion in these passages to show "that [Korsgaard] judges agents to be unified to the extent that they exhibit a kind of [efficacy]."[14] But as we saw with (c) above, Katsafanas clearly thinks that it is Korsgaard's view that the *CI* is the only principle the enables full efficacy. But this is false. For Korsgaard, the *moral law* is the only principle that, *in itself* enables full efficacy. And, for her, an action is good only if its principle is the *moral law*.

[14] See, for example, p. 93 where Katsafanas cites *Self-Constitution* (pp. 162-3, 179, and 174).

Katsafanas, then, seems to be eliding the distinction Korsgaard draws between the CI and the moral law.

Whatever the case, one might think that this leaves Korsgaard no better off, for replacing the CI with the moral law makes conformity to the moral law a sufficient and necessary condition for action:

> (a*) An agent's A-ing is an action iff A-ing is attributable to the agent as a unified whole.
> (b*) An action is attributable to the agent as a unified whole iff the principle of the agent's action renders the agent efficacious.
> (c*) The **moral law** is the only principle which renders the agent efficacious.
> (d*) Therefore, an agent's A-ing is an action iff the principle of the agent's A-ing is **the moral law**.

But (c*) is false. The moral law is the only principle which *as such* renders the agent efficacious; but as we've seen, other principles can render agents efficacious too, it is just that those principles require contingencies we have no control over to obtain. This means we should switch out (c*) with the following:

> (c**) the **moral law** is the only principle which *as such* renders the agent efficacious.

But now (d*) no longer follows. As long as the agents acts from some general principle – e.g., the principle of desire satisfaction – the agent can be efficacious and what she does can count as an action. But in such a case the agent in question would be in conformity with the CI but not the ML.

While we've lifted ourselves out from under the first horn of Katsafanas' dilemma, doing so leads us right into the second horn: if acting requires conformity to the CI but not conformity to the moral law, then it would appear that I am not *committed* to the moral law simply in virtue of acting; morality, it would seem, is optional.

7. Responding to Katsafanas' Dilemma: Horn 2

For Korsgaard, action's dual functions of efficacy and self-constitution turn out to be interdependent commitments. In acting I commit myself to being efficacious in the pursuit of an end, for to be otherwise would involve giving up the end, and so to give up acting. But now, I am not efficacious if a desired result simply happens. In order to be efficacious, the desired result must be one that *I* produce. And in order for it to be a result that *I* produce, the performance that brings the result about must be attributable to me *as a whole*. This requires that I be a unified agent. So my commitment to efficacy *ipso facto* also commits me to constituting myself.

As we've seen, I am made a constituted whole by the principle on which I act. But not all principles are equally conducive to self-constitution. Rather, the only principles which, according to Korsgaard, *guarantee* self-constitution are those that conform to the moral law. But this means that, in being committed to efficacy, I am also committed to acting on principles that conform to the moral law. We arrive at the conclusion we needed in order to get out from under the second horn: our commitment to efficacy in the pursuit of our ends commits us to acting on principles that conform to the moral law.

Here is a more formal representation of the argument just sketched:

1. In acting I am committed to efficacy in the pursuit of my ends.
2. In order to be efficacious, I must be a constituted whole.
3. I am made a constituted whole by the principles on which I act.
4. Constitution as a unified self is *guaranteed* only by acting on principles that conform to the moral law
5. So, in being committed to efficacy I am also committed to acting on principles that conform to the ML.

How does this show that the moral law is not, in fact optional? Recall the second horn of Katsafanas' dilemma: if it is possible to act without *fully* constituting ourselves, that is without conforming to the moral law, then acting *as such* does not commit us to fully constituting ourselves, and so does not *commit* us to conforming to the moral law, but rather conformity only to a thinner principle. The solution to the problem lies in recognizing that in the first instance, acting commits one to efficacy, and that it is in virtue of *that* commitment that I then, in the second instance, commit myself to the moral law (since that is the only principle that guarantees my efficacy).

This solution leaves us with plenty of room for bad actions. My *commitment* to efficacy *commits* me to the moral law, but the distinguishing mark of action is not *conformity* to the moral law. It is often not obvious which principles are consistent with the moral law, nor that my commitment to efficacy commits me to the moral law (after all, it took a rather dense philosophical argument to bring that fact out). So we can choose a principle, falsely thinking that it is conducive to self-constitution and efficacy, or we can simply be ignorant about the structure of our agency. While it would appear this means that ignorance of some kind will always appear in the etiology of immoral action, it's important to note that we are not left with a Socratic guise of the good thesis. It is fully consistent with this view that I can act in full knowledge that my action conflicts with the moral law. What follows, however, is that I can only act in such a light if I am ignorant of how so acting will threaten to undermine my efficacy.

8. Are We Really Committed to Performing Good Actions?

One might still be skeptical that we are committed to performing a good action simply in virtue of being commited to acting. Recall the example of the house builder introduced earlier. The example is meant to show that, just as a house builder is committed to building a good house simply in virtue of building a house, so is an agent committed to performing a good action simply in virtue of performing an action. Barandalla and Ridge (2011) have recently argued that the analogy doesn't work: there is little reason to think that a house-builder is committed to building a good house simply in virtue of building a house.

> Suppose the [house] builder knows for sure that his customers can tell if what he has built qualifies as a house at all. In that case, it will be important to him that what he builds does qualify as a house, as otherwise he will not be paid. … So the devious builder aims to build a house which is just good enough, in terms of the perfect ideal of a house, to count as a house, but no better. That seems like a perfectly coherent possibility, but if it is then Korsgaard can no longer use this example to show that one cannot, after all, literally choose a bad action. (pp. 378-379)

If we individuate functional kinds according to *function* (rather than form), and insofar as artifacts gain their function from the intentions of the designer, then an object is a house when one designs it with the intention that it *serve as shelter*. It follows that, in order to intend to build a house, one must intend to build something that can serve as shelter.

But what is it for a functional item to be *good* as an instance of its kind? There are two ways to answer. First, a functional item can be said to be good if it fulfills its function. A stapler that staples paper is a good stapler; one that doesn't staple paper is a bad stapler. The second way a functional item can be good is if it not only fulfills its

function, but does so excellently.[15] A good stapler, in this sense, would be one that can staple large amounts of paper, that doesn't require a lot of force, and so on. When Korsgaard says that the house builder is committed to building a good house, which sense of 'good' does she have in mind?

If Korsgaard has the first sense of 'good' in mind, then a good house would simply be a house that serves as shelter. If this is all she means, then, as we've seen, it is rather straightforward that a house builder must intend to build a "good" house. However, mere function fulfillment is a fairly minimal good, and our ordinary uses of 'good' rarely track the mere fulfilling of a function.

Maybe Korsgaard has the second sense of 'goodness' in mind. A house would be good in this sense if, perhaps, it provides a shelter that is *comfortable*. But why think that a builder is committed to building a house that functions *excellently* simply in virtue of intending to build a house? One possibility is that artifacts that do not function *excellently* will have a tendency to undermine the object's very ability to fulfill its function. A house builder must intend to build a house good along the second dimension, then, because if she doesn't, she'll be trying to build something that might very well not serve as shelter, which would contradict her intention.

While it may be plausible that in some instances non-excellence in functioning may undermine an object's ability to fulfill its function – we can easily imagine this

[15] It's not clear to me how to flesh out this second sense of 'good'. One option would be to tie these goods to some value external to the nature of the thing, such as an aesthetic value. I think this clearly can't be Korsgaard's view as the constitutivist project is to ground norms in the nature of the object or activity the norm is governing. This suggests that this second sense of good we are considering must in some way be tied to the object's *function* and *form*. The suggestion I am considering in the main body is that an object X is good in this second sense of good if it performs its function F *excellently*.

being the case with a stapler that is difficult to operate – it's unlikely that this will be a universal feature of teleologically organized objects, or even a common one. It certainly doesn't seem true of houses.

In the end, I think the house example just does not work. But this doesn't undermine Korsgaard's view. What is important for the view is not the claim that in order to act we must *intend* to perform a good action, but rather the claim that in acting we are *committed* to performing good actions. Barandalla and Ridge nevertheless think Korsgaard has not shown even this:

> [One] can intend to act in a way that will constitute him as an agent, but he [need not be] committed to constituting himself as a *good* agent any more than our devious builder is committed to building a *good house*. (pp. 379)

Barandalla and Ridge recognize that Korsgaard has a response: an agent who is not committed to fully constituting himself would be making a merely *conditional* commitment to his agency. Such an agent would be committed to efficacy but only on the condition that certain external contingencies obtain. But one can't do that. Barandalla and Ridge are not impressed with this response:

> It is unclear even on her own account *why* someone cannot make a merely conditional commitment to his own agency. Korsgaard asserts this as an obvious consequence of her view, but we do not see how it follows from what she has established... (ibid)

Here is why it is a consequence of Korsgaard's view. To have a commitment to one's own agency is to be committed to efficacy in the pursuit of one's ends. To be conditionally committed to one's own agency, then, is to be conditionally committed to efficacy in the pursuit of one's ends; that is, to be committed to efficacy but only if certain conditions are met. Recall that the efficacy of one's agency is a product of the

principles that one adopts. The "lesser" principles unify one's agency, and so make one efficacious, but only in the presence of contingencies beyond one's direct control. I am conditionally committed to my own agency, then, if my commitment is conditional on these contingencies obtaining. But if I've adopted such a principle knowing *full well* that my efficacy is dependent on the realization of these contingencies, then I have, in effect, given up on my commitment to efficacy and so abdicated control of my agency.

9. Completing the Solution

We are not out from under the bad action problem quite yet. Recall that, for Korsgaard, the form of a house is that which *enables* it to fulfill its function, namely to serve as shelter. If an object *lacks* a form conducive to shelter—that is if it cannot serve as shelter—then it is not a house. As we saw in section 7, this idea leads to a problem, for the function of *action* is efficacy and self-constitution. The form of an action, then, must be that which *enables* the agent to be efficacious and constituted. If a behavior lacks this form, it would seem that it is not an action. Now, the form of an action is the principle behind the action. But efficacy and self-constitution—action's two functions—would seem to truly only be obtained when one acts according to the moral law. So, the *moral law*, it would seem, is the form of action, for this is the principle that *makes* an agent efficacious and autonomous in any circumstance. But if the moral law is the *form* of action, and we individuate according to form, then in order for a behavior to count as action it must conform to the moral law. Now it looks like there can be no bad actions.

Before I walk us through the solution, I'd like to point out that the argument just sketched is similar in many ways to one made by Silverstein (2016):

> [Korsgaard's] Aristotelian conception of action's constitutive standard leaves no room for the possibility of bad actions that run afoul of this standard. Activity through which an agent does not constitute himself is not *bad* action. Since it fails to satisfy action's constitutive norm, it also fails to satisfy action's constitutive condition. Consequently, *it is not action at all*. (p. 217)

Silverstein continues:

> The problem is not specific to the particular constitutive norm Korsgaard identifies (namely self-constitution). It arises instead from her general conception of constitutive standards as Aristotelian forms. On this conception, "every object and activity is defined by certain standards that are both *constitutive of it and normative for it.*"[16] There is no room for defective actions because what determines whether something is an action (its function) is *the very same standard* as what determines whether it is a good action (its constitutive norm). The same goes for houses and axes – and for anything else governed by this sort of constitutive standard.[17] (p. 219)

Like Katsafanas, Silverstein clearly thinks that Korsgaard suffers from the bad action problem. However, Silverstein identifies the source of her bad action problem squarely in her teleology. Earlier in Silverstein's paper, he states that for Aristotle and Korsgaard, "membership in a category or kind is determined by the candidate member's form or essence". He goes on to add that, "a thing's form is determined by its purpose or

---

[16] Silverstein cites Korsgaard (2009: 25).

[17] Silverstein offers the following footnote here:

> Korsgaard's discussion of the categorical imperative confirms the assessment of her view. She claims an agent who fails to will his maxim as a universal law would have to engage in "particularistic willing." But particularistic willing is impossible, according to Korsgaard… Thus, the only alternative to acting in accordance with the categorical imperative is not acting at all. There is no such thing as genuine action that violates the categorical imperative. We cannot act badly. (2016, p. 6, fn. 11)

Here Silverstein is making a similar mistake to the one I accuse Katsafanas of making in section 6. My response to Silverstein is the same as my response to Katsafanas: conformity to the CI is constitutive of action, but conformity to the CI (as such) does not a good action make. But one need not fall prey to this error in order for the current worry to apply to Korsgaard.

function." Silverstein recognizes the same problem I do with Korsgaard's teleology: "an axe that cannot chop is not a bad or defective axe. Instead, it is not an axe at all."

The individuation principle I propose above provides us with a solution. If we separate form from function and individuate according to the latter we can make room for bad action. An action is bad when its form – its principle – is not conducive to efficacy and self-constitution. But even when a principle *fails* to support efficacy and self-constitution, we can still say that the behavior was an action, at least insofar as the behavior has the function of seeking efficacy and self-constitution. And, as I suggested in section 7, behavior acquires this function be virtue of it's relation to an agent's intentions.

Silverstein is cognizant of the strategy I propose; however, he thinks it is not available to Korsgaard for three reasons. First, it violates a "central metanormative commitment" of Korsgaard's, namely the identification of the real with the good. Second, Silverstein suggests that the teleology I propose would impose *external* standards, not *internal* ones. Finally, the view is inconsistent with Korsgaard's claim that agents constitute themselves by acting.

Let's deal with each of these objections in turn. First, it's not clear to me how we should go about understanding what, exactly, counts as Korsgaard's "central metanormative commitments". One of Korsgaard's metanormative commitments concerns grounding normativity in the teleological structure of the will. This level of stating the commitment leaves it open as to how we are to understand the teleology. I've been arguing that her theory of teleology has weaknesses, but it can be replaced with a better one without costing anything of consequence. And since we can do this, we should.

This line of thought brings us to Silverstein's second point: I am imposing *external* standards on action, rather than *internal* ones. A standard, I take it, is *internal* to an object or practice if the standard is rooted in the *nature* of the object or practice itself. By individuating according to function rather than form, we do not impose external standards on the object. This is because the object's *kind* is determined by its function.

Why does Silverstein insist that this strategy involves imposing an external standard? He seems to think that, particularly in the case of artifacts, because functions would be derived from the intentions of the designer or user, this shows that the functions, and thus the standards they give rise to, are actually *external* to the object itself—they reside in a person's mind.[18]

I think that Silverstein imposes an external/internal dichotomy where one doesn't belong. An object's *kind* does not inhere in the physical substrate of the object itself. Kinds are types, individual objects are tokens. The view under consideration holds that an individual is a token of a functional type in virtue of that's object having a particular function. While it's true that with artifacts, an object's function lies in the mind of a designer or user, we shouldn't lose sight of the fact that artifacts are extensions of human activities. Objects *become* artifacts through our *using* them. Their natures *as artifacts* are very much tied up with our intentions. An artifactual object becomes a token of a

[18] He does suggest in passing that "Biological kinds such as eyes and hearts might also be example of functional kinds in this sense" (p. 222). While Silverstein doesn't identify the external source of an eye's function, one might reasonably think that natural selection is to the function of an eye what a designer's intention is to the function of an axe.

functional kind by being wrapped up with an agent's activity. We cannot separate the nature of functional objects from the mind of the designers or users.[19]

Silverstein's final objection, recall, is that such a view would be inconsistent with Korsgaard's claim that agents constitute themselves by acting, for "[i]f merely attempting to constitute oneself were enough to satisfy agency's constitutive condition, then one could meet that condition while failing in one's attempt at self-constitution. In such cases, one both would and would not be a unified self or agent" (p. 224). According to the view under consideration, a behavior counts as an action if the behavior has action's function, namely self-constitution and efficacy. A behavior has such a function if, in behaving, the agent makes a commitment to being an efficacious cause. The way one makes a commitment, according to Korsgaard, is by adopting a principle. Whenever one exercises one's will in this way, one acts. But this doesn't mean that one's action is successful, in the sense of achieving efficacy and self-constitution. By making a commitment to being an efficacious cause, I perform an action whenever I will, for my behavior, through this commitment, has the function of efficacy and self-constitution. This doesn't impugn the claim that we constitute ourselves by acting, though it does impugn the claim, at least on the surface, that we can only act when we are self-constituted. What this does suggest, though, is that we need to be more careful in distinguishing the *performance* of acting from *being efficacious* in said performance. One acts whenever one wills; but willing is not always efficacious.

[19] Korsgaard makes a similar point (2009, pp. 39-40).

10. Implications for Constitutivism

This essay has argued, against many critics, that Korsgaard's view need not give rise to *the bad action problem*. At this point I'd like to draw the reader's attention to a particular feature of the discussion that, I think, has interesting implications for how we should think about constitutivism.

I have argued that on Korsgaard's view morality is best understood as arising out of a series of commitments, some of which are constitutive of agency *as such*, some of which are constitutive of the *kinds* of agents we are. Consider Table 1, which lists the different commitments, their content, and their source.

Table 1

Commitment	Content	Source
Efficacy (HI)	Cause a change in the world	Our nature as agents *qua* agents.
Choosing a principle (CI)	Act on a principle; any principle will do.	Our nature as *intentional* agents.
Moral Law	Act on a principle that can be *willed* as a universal law for everyone.	Our nature as reflective, temporally extended social agents.

On Korsgaard's view, our commitments to efficacy and the CI are grounded in features of our agency that we share with other animals.[20] Our commitment to the moral law, on the other hand, is another matter. Non-human animals are not bound by the moral law. The reason for this has to do with features that are particular to *our agency*: our capacity

[20] In Chapter 5 of *Self-Constitution,* Korsgaard's argues that animals, much like humans, need to conform to the CI in order to act. The difference with animals is that for them it is much easier, since *instinct* provides them with a principle. In this sense, instinct is the analog to reasons in intentional agency.

for reflection, the fact that our agency is temporally extended, and the fact that we are *social agents,* that is, agents who can *share their agency.*[21]

When we understand the underlying structure of Korsgaard's view, we can see a path forward for one who is friendly toward constitutivism but finds the metaphysics of self-constitution unappealing. The key to making constitutivism work, I'd like to suggest, lies in the idea that agency is structured by *commitments*; what distinguishes different kinds of agents is the different kinds of commitments that structure their will. The key question for the constitutivist, then, should *not* be what distinguishes action from mere behavior, but rather what is distinctive about human agency, that is, what distinguishes our agency from all the other kinds of agency we've encountered on Earth. While Korsgaard's work strongly suggests that this difference lies in our powers of reflection which create a gap we have to bridge, it is not until we include our nature as *social agents* – perhaps another distinguishing feature of our will – that we actually arrive at the conclusion that we are committed to the moral law.

I think that Korsgaard has identified a deep truth here about our nature as agents – overlooked by friends and critics of constitutivism alike – a truth that gets obscured by the metaphysics of self-constitution. Might the key to constitutivism be not the metaphysics of self-constitution, the desire for self-understanding, or the aim of will-to-power, but rather a commitment to *shared agency* that is grounded in our nature as social

[21] Many constitutivists hope to explain the universality and categoricity of morality by appealing to features of our agency that distinguish action from mere behavior. Korsgaard's work shows that this is insufficient. A constitutivist also needs to appeal to features of our agency that distinguish us from the other animals of the Earth, lest our theory lead to the absurd conclusion that monkeys, mules, and mice have moral duties.

agents? I think the answer to this question is yes. I will turn to defending this claim in the next chapter.

11. Conclusion

In Chapter 1, I suggested that constitutivism is motivated by what I called *the core analogy*: morality is to agency what the rules of chess are to chess. How does understanding agency as structured by commitments relate back to the core analogy? In chapter 2 I suggested that much as *agency* is structured by commitments, we can understand a lot of *activities* as being structured by commitments. Morality is to agency as the rules of chess are to chess insofar as both of these are commitment-structured activities, and a commitment structured activity is a norm-governed activity. We escape the bad action problem, however, when we recognize that our agency has a far more complicated structure than a game. Success in performing an action requires only efficacy in the pursuit of our ends – or, as I put it in chapter 2 – conformity to the instrumental principle. But, as I emphasized in section 10 of this chapter, this does not exhaust the commitments that structure *our* agency, for we are also committed to *efficacious social interaction*. If the moral law is a constitutive principle of that activity, then a commitment to complying with the moral law would follow.

Chapter 4: Is the Moral Law a Constitutive Principle of Shared Agency?

At the end of Chapter 3 I suggested that Korsgaard identifies the moral law as a constitutive principle of shared agency. Let's call this the **Shared Agency Hypothesis**:

> **Shared Agency Hypothesis:** The moral law is a constitutive principle of shared agency in that, in order for two or more parties to share their agency they must treat each other as ends-in-themselves.

The aim of this chapter is to evaluate Korsgaard's argument for this provocative thesis and determine if it has any merit. I will argue that Korsgaard's defense of the Shared Agency Hypothesis fails, but that a defense of it can nevertheless be offered.

I have split this chapter into two parts. Part I interrogates Korsgaard's defense of the Shared Agency Hypothesis. Doing so requires making sense of some of the toughest pages in Korsgaard's oeuvre, namely those where she discusses the nature of so-called "public reasons". If this chapter does nothing else, hopefully it will at least help us understand what she is up to in those pages.

In Part II, I turn to my own argument in favor of the Shared Agency Hypothesis. This argument is rather complicated. It will involve explicating the content of the moral law (i.e., what it means to treat another as a mere means) and investigating two major theories of shared agency. What we will find is that shared agency has constitutive norms akin to the instrumental principle of singular agency.[1] But unlike the norms of singular agency, these norms are *interpersonal* and *directional*: they come in the form of

[1] Following Margaret Gilbert (e.g., 2009), I will use the phrase 'singular agency' to refer to the agency of an individual in contrast to the agency of a group or duo, i.e., the kind agency that is shared.

obligations and corresponding entitlements. Shared agency, then, is a kind of *normative relationship*. I will argue that the moral law describes a condition on the possibility of parties entering into this normative relationship. In particular, I will argue that if person A were to use person B as a mere means in order to enter into this normative relationship, A would in effect unilaterally obligate B to A. I will argue that because this is not possible, the moral law must be a constitutive principle of shared agency.

But I get ahead of myself. Let's turn to the investigation of Korsgaard's defense of the Shared Agency Hypothesis.

Part I: Korsgaard's Derivation of the Moral Law

1. Korsgaard's Collective Action Argument

Korsgaard defends the claim that the Moral Law is a constitutive principle of shared agency in Chapter 9 of *Self-Constitution*. This chapter is a complex and sometimes meandering discussion of the nature of what she calls "interaction". My aim in this section will not be to summarize all aspects of her discussion, but rather to interpret the core of the argument.

Korsgaard's argument is motivated by a puzzle about interaction found in Kant's *Lectures on Ethics*. The puzzle concerns what we might call, tongue-in-cheek, "the problem of sex". How can we interact sexually with a partner without treating them as a mere means, whether it be an instrument for sexual satisfaction, or an object of aesthetic enjoyment? Kant thought the desire inherent in sex is one of possession, so to desire

someone sexually would involve the desire to possess someone as a mere object.[2] To satisfy such a desire would require *possession* of a person. But people are not property to be owned. So sex, it would seem, cannot be a form of interaction in the Kingdom of Ends. What a boring place!

As one would expect from a man of his day, Kant's "solution" to this problem is marriage,[3] which (somehow) involves the unification of wills. When our wills are unified, my possession of you is equivalent to you possessing yourself, and vice-versa. In the Kingdom of Ends, sex is an interaction that only married individuals undergo. What a boring place!

For Kant, sex is not the only puzzle which a unification of wills solves. How can we devote ourselves to the happiness of someone else without turning ourselves into a doormat to be trampled on? The solution to this problem is friendship. In a friendship, two people unify their wills, each looking after the other's happiness. If the wills are unified in this way, one doesn't become a doormat when devoting oneself to the happiness of another since, in doing so, one's own happiness is being cared for.

Korsgaard believes that sex and friendship are not unique forms of interaction. Rather, she thinks that "everyday interaction" also requires the unification of wills that we see in these more special cases. This raises a few questions. First, what does Korsgaard mean by "everyday interaction" and why does it require a unification of wills? Second, what does she *mean* by "unification of wills" and how, exactly, do wills become

[2] Kant was a product of his time.

[3] Of course, the *real* solution to this problem lies in reconceptualizing the masculine attitude of sex. This, however, is a different dissertation.

unified? Finally, what is the puzzle that everyday interaction poses which the unification of wills supposedly solves?

Let's start with the first of these questions: what does Korsgaard mean by "everyday interaction"? The word "interaction" can mean several things. It can mean two individuals *coming into contact* with each other, such as how particles "interact". It can refer to individuals *responding* to each other, such as when I swerve to the right in my car when a car to my left inches into my lane. When Korsgaard uses the word "interaction" she means neither of these two things. Consider the following passage:

> When we interact with each other what we do is deliberate *together*, to arrive at a **shared decision**. Since the conclusion of a practical syllogism is an action, **the result is an action that we perform together**, governed by a law we freely choose together. (2009, p. 190)

What Korsgaard means by "interaction" is what most other theorists mean by "collective action", "shared action", or "group action".

The puzzle which Korsgaard thinks lies at the heart of collective action is the same sort of puzzle which, in the previous chapter, we saw she thinks lies at the heart of *singular action*. On Korsgaard's view, in order for an individual to act she must bridge the gap created by reflection. This requires adopting a principle capable of adjudicating between competing inclinations. This process of bridging the gap is what Korsgaard thinks we refer to with the phrase "practical reasoning". The result of this process of reasoning is an action, the preceding event of which is the unification of the individual's will (this is, I take it, is what bridging the gap amounts to). Now, in the case of collective action we have a more complicated picture, for we have several different wills each of

which has to be unified – and then what? We know that an action is the end result, but what precedes the action?

In the case of singular action, we saw that what precedes an action is practical reasoning, which is a process by which the individual's will is unified. From the passage quoted above we can infer that Korsgaard thinks collective action is proceeded by *shared deliberation*, which we might think of as the collective analog of practical reasoning. The analog for collective action of the gap that exists between inclination and action for singular agents is what we might call "the spatial gap" that exists between the wills of different agents. Much as singular agency requires adjudicating between competing inclinations, collective action requires adjudicating between potentially competing wills. The solution to this competition, much as in the case of singular agency, is the unification of these disparate wills through the process of practical reasoning, in this case collective reasoning.

To summarize: For Korsgaard, "everyday interaction" is what is often referred to as "collective action" or "shared agency". The puzzle behind shared agency concerns how a collective action could result from a collection of disparate wills. The solution to this puzzle – according to Korsgaard – is that shared deliberation results in a unification of wills. This, according to Korsgaard, is what makes collection action possible.

So collective action requires that parties to the collective act unify their will. But where does morality come into the picture? We've seen that for Korsgaard, to perform a collective action parties to the action must unify their will. In the previous chapter we saw that with cases of singular agency an individual need only adopt a principle (and

have some luck) to do so. Why isn't it the same with collective agency – why is conformity with the moral law in this case *necessary* for unity? Korsgaard supplies the answer to this question in two steps. Consider step 1:

> The possibility of personal interaction depends on the possibility of shared deliberation. And that possibility in turn depends on a certain conception of reasons. Our reasons must be what I call public reasons, whose normative force can extend across the boundaries between people.[4] (2009, p. 191)

Here Korsgaard is claiming that unifying our will through shared deliberation requires that participants to a collective action act on what Korsgaard calls "public reasons".[5] We get to morality in the second step:

> I just claimed that if personal interaction is to be possible, we must reason together, and this means that I must treat your reasons … as public reasons. And to the extent that I must do that, I must also treat you as what Kant called an end in yourself. (2009, p.192)

For Korsgaard, a condition on my acting on public reasons in my interactions with you is that I must treat you as an end in yourself and this is nothing less than acting on the moral law. So on Korsgaard's view treating you as an end is a condition on the possibility of engaging with you in shared deliberation, which in turn is a requirement for us to unify our wills in an effort to act collectively.

We can represent this argument more straightforwardly as so:

1. Collective action requires shared deliberation.
2. Shared deliberation requires acting on public reason.
3. Acting on public reason requires complying with the moral law.
4. So, collective action requires complying with the moral law.

---

[4] Here Korsgaard footnotes her famous discussion of public reasons in *Sources of Normativity*, sometimes called "The Private Reasons Argument".

[5] We will interrogate what Korsgaard might mean by public reason in a bit.

Let's call this argument the *Collective Action Argument*. The concept of a public reason plays a central role in moving from the uncontroversial first premise, to the controversial conclusion that the moral law is a constitutive principle of collective action. I turn now to an interrogation of this concept.

2. Korsgaard on Public Reason

We've seen that the concept of public reason plays an important role in Korsgaard's *Collective Action Argument*. In order to act collectively, Korsgaard argues we need to deliberate collectively. But collective deliberation requires "acting on public reason" and doing this requires complying with the moral law. But what is it to act on public reason, why does shared deliberation require doing so, and what does this have to do with the moral law? In this section I will attempt to answer these questions. We will discover that the concept of public reason is not well-defined, and that Korsgaard in fact equivocates when she uses it.

Before we proceed, it is important not to confuse Korsgaard's concept of a public reason with John Rawls' somewhat related concept which goes by the same name. In Rawls' political philosophy, a public reason is a reason one can permissibly invoke to justify legislation and judicial decisions. Public reasons, so understood, are reasons that do not make reference to controversial comprehensive doctrines, such as a particular moral theory, religion, or philosophical worldview. Public reasons are supposed to be

reasons that all citizens "share" *regardless* of what our comprehensive doctrines happen to be.[6]

Korsgaard certainly uses the phrase "public reason" because of Rawls' use of it in his political philosophy; however, in Korsgaard's hands the concept is not political, but rather metaphysical. It describes either a *kind* of reason or a certain *character* that reasons have. When Korsgaard is discussing public reasons, she is discussing the *nature* of practical reasons.

So what *is* a public reason for Korsgaard? It turns out the answer to this question is not so straightforward. She discusses the concept in three different texts (1996a, 1996b, 2009). The answer to this question will depend on which of these three texts you are engaging with. Her most famous discussion of public reasons occurs in the so called "Private Reason Argument" of *Sources of Normativity*. In *Self-Constitution* she moves away from some of the central claims she makes there regarding public reasons (or so I will argue). The issue of public reasons comes up– though without using that phrase – in "The Reasons that We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Value", an essay that predates *Sources*. Part of the aim of this section will be to decipher what she is talking about in these texts. But the main task is to identify what her considered account of public reason is in *Self-Constitution*, since it is in this text that we find the Collective Action Argument.

[6] The canonical statement by Rawls about public reason occurs in Rawls (2005).

Our inquiry into the concept of public reason begins with Korsgaard's initial gloss in *Self-Constitution*. She states that a public reason is a reason "whose normative force can extend across the boundaries between people" (2009, p. 191). She continues:

> Public reasons are roughly the same as what are sometimes called objective or agent neutral reasons. They may be contrasted to what I call private reasons – subjective or agent relative reasons. A private reason is a reason whose normative force is private, in the sense it belongs to only one person. (ibid)

It might seem that there is just one definition of a public reason being offered here, but there is actually two. They perhaps overlap but are not equivalent. According to the first definition, a public reason is a reason whose normative force is shared, which contrasts with a private reason whose normative force is not. According to the second definition, a public reason is the same thing as an agent-neutral reason.

I said that these two definitions are not equivalent. Why is that? Let's start with the idea of an agent-neutral reason. The distinction between an agent-neutral and agent-relative reason is due to Thomas Nagel (1970, 1986) and Derek Parfit (1984). The distinction, strictly speaking, is originally Nagel's; the nomenclature is Parfit's (1984).

In the *Possibility of Altruism*, Nagel draws a distinction between subjective and objective reasons. His articulation of this distinction is a bit cumbersome, but technically precise and helpful once it is thoroughly unpacked (despite the unhelpful terms 'subjective' and 'objective').

To begin, let's consider Nagel's definition of a reason: "Every reason is a predicate R such that for all persons p and events A, if R is true of A, then p has *prima*

facie reason to promote A" (1970, p. 47). Here is how Nagel would represent this formally:

(p, A) (If R is true of A, then p has reason to promote A).

Let's call this a "reason statement". Nagel notes that "the predicate R may, or may not, contain a free occurrence of the variable p" (p. 90). Nagel calls such a variable a "free-agent variable", and the reasons that contain free agent-variables subjective reasons. All reasons and principles expressible in terms of a reason statement either contain a free agent-variable or they do not. The former Nagel calls subjective; the latter will he calls objective.

Nagel suggests that the following example will help clarify:

> Suppose G.E. Moore finds himself in the path of an oncoming truck, and concludes that he has reason to remove himself. ... If he is asked what reason he has to get out of the way, he may say (among other things) any of the following
>
> (a) that the act will prolong G.E. Moore's life;
> (b) that the act will prolong his life;
> (c) that the act will prolong someone's life. (ibid, p. 91)

Nagel says that only (b) is a subjective reason. Here are the corresponding formalizations:

(a*) (p, A) (If A will prolong Moore's life, then p has a reason to promote A.)
(b*) (p, A) (If A will prolong p's life, then p has a reason to promote A.)
(c*) (p, A) (If (if (∃q) (A will prolong q's life), then p has a reason to promote A.)

It might not be obvious why (b*) is the only reason statement with a free-agent variable. Aren't the variables bound in all of these sentences? The unclarity here is due to Nagel's

presentation. It is important to keep in mind that the predicate R is that part of the antecedent which applies to A (in bold):

(b*) (p, A) (If A **will prolong p's life**, then p has a reason to promote A.)

While the variable in the antecedent is bound by the universal quantifier, it is not bound *within* R. Compare (b*) with (c*). In (c*) the variable in R – namely 'q' – is bound by the existential quantifier in the antecedent. In (a*) there is a rigid designator, but no variable.

While Nagel's distinction between subjective and objective reasons is precise, so articulated there is not much intuitive force behind it. What exactly is being captured in the distinction? The closest we come to an intuitive statement of what the formalism captures is when Nagel sates that "objective reasons... contain no open reference to the doer of any act to which they may be applied" (ibid). The intuitive idea behind the distinction, then, is that subjective reasons have within their scope a particular agent, and this agent is the one who will be performing the act in question. So subjective reasons *make essential reference to the individual who will be potentially acting on the reason*. Objective reasons make no such reference, and as such count as considerations for any person who finds themselves able to act.[7]

Parfit (1984), helpfully offers a gloss on the intuition driving Nagel's distinction and in so doing introduces the now canonical 'agent-neutral/agent-relative distinction':

> Nagel calls a reason objective if it is not tied down to any point of view. Suppose we claim that there is a reason to relieve some person's suffering. This reason is objective if it is a reason for everyone – for anyone who could relieve this

[7] Consider, for example, the reason statement "The act will prolong GE Moore's life". This is, presumably, equally a reason for you, or me, or anyone just as much as it is for GE Moore to get GE Moore out of the way.

> person's suffering. I call such reasons agent-neutral. Nagel's subjective reasons are reasons only for the agent. I call these agent-relative. (p. 143)

When a reason makes an essential reference to the agent performing the act, such reasons are "tied down" to a particular "point of view", namely that of the agent acting. The reason I have to submit a chapter of my dissertation for publication is agent-relative; it is essentially tied to my point of view. The reason I have to vote Democratic in the 2018 midterms is agent-neutral; it is not tied to my point of view, but rather to the health of our country. It is a reason that everyone has. (Even, I assume, registered Republicans!)

Insofar as agent-neutral reasons are reasons *not* tied to any particular point of view, it would seem that this tracks closely with what Korsgaard is calling a "public reason", at least if we are understanding such reasons as those whose normative force is shared. And agent-relative reasons, in being tied to a particular point of view, would seem to track "private reasons", reasons whose normative force is felt only by the person whose reason it is.

I just said that the public/private distinction might indeed seem to track the AN/AR distinction. But earlier I said these are not equivalent. Which is it? The answer depends on what it means for the normative force of a reason to be shared.[8] To see the problem here it will be helpful to turn back to earlier work in which the concept of public reason plays an important role: *The Sources of Normativity*.

[8] What is the normative force of a reason? This is not a phrase that is often parsed in the contemporary philosophical cannon. I take it that a consideration or fact R has normative force for person P just in case R makes a pro tanto rational claim on P. This way of cashing out the phrase is unfortunate, in that it trades one obscure metaphor for another.

In her "Private Reasons Argument", Korsgaard aims to show that all reasons are public and that there are no private reasons.[9] Her main argument for this claim is modeled after Wittgenstein's famous "private language argument" from *Philosophical Investigations*. Reasons by their very nature are normative. Insofar as reasons are normative, they must set a standard of success and give rise to a possibility of error. For example, if the weight of reasons tells me to donate to Oxfam, this establishes a standard of success for my actions. I succeed if I donate, fail if I don't. By providing me with a standard, they operate as a guide.[10] Now, In order for reasons to establish a standard of success and thus create a possibility of error, there must be a rule that stipulates when success is achieved and when error occurs. This is where the private language argument enters the picture. Rules, by their very nature, are public. It follows that reasons by their very nature must also be public. So all reasons are public reasons and there can be no private reasons.

But this argument only works if a public reason is public in the sense that *Wittgenstein* meant. Recall Korsgaard's stipulation that public reasons are equivalent to agent-neutral reasons. Does the private reasons argument establish *that*?

Joshua Gert (2002) has shown, I think convincingly, that the answer to this question is no. Recall that a public reason, by Korsgaard's own stipulation (SN 133, fn 3) is supposed to be identical to an agent-neutral reason in Nagel's and Parfit's senses: a

---

[9] As she does in *Self-Constitution*, Korsgaard states in *Sources* that the public/private distinction maps the AR/AN distinction (see 1996a, p. 133, fn. 3).

[10] On the idea that a standard can be normative only if it makes error possible, see Korsgaard's "The Normativity of Instrumental Reason" in her (2008)

reason that does not make essential reference to the person whose reason it is. A private reason, in contrast, is a reason that *does* make essential reference to the person whose reason it is (such as my reason to play with *my* cat). Compare this with Wittgenstein's understanding of "public" and "private" in the Private Language Argument. According to Wittgenstein, a private language is one which, in principle, only the speaker can understand because "the words of this language are to refer to what can be known only to the speaker," while a public language is one which is, in principle, understandable by all (PI §243). As Gert points out, agent-relative reasons, in making essential reference to the agent acting on the reason, operate like indexicals, which make essential reference to the speaker, space, or time of utterance. But indexicals are just as public in Wittgenstein's sense as non-indexicals, for indexicals have public rules by which we can judge correct usage. (If they didn't, we couldn't learn how to use them correctly, and there couldn't be correct or incorrect usage.) What Korsgaard's Private Reasons Argument shows, then, is that all reasons are public in Wittgenstein's sense, namely that the normative structure of all reasons is communicable. If I have a reason to till my garden, I can (in principle) communicate to you the underlying rule that makes this so. But this doesn't imply that all reasons are agent-neutral.[11]

We are now in a position to understand my claim that in *Self-Constitution* Korsgaard offers two distinct glosses on what a public reason is. One gloss was that they are agent-neutral reasons, the other was that they are shareable reasons. If what it is for a

[11] Once we recognize what is going on here, we can see that the result of the private reason argument is unsurprising. Wittgenstein's argument, on the other hand, was perhaps surprising because it showed that the language was essentially normative, which is perhaps a surprising thing to discover. But this is not a surprising thing to discover about reasons, which are normative if *anything is*.

reason to be shareable is identical to what it is for a language to be public – this seems to be the claim in *Sources* – then these two different glosses are in fact distinct (even if unintentionally so), for as we've seen publicity is not the same as agent-neutrality. But in *Self-Constitution* Korsgaard seems to have moved away from the claim that all reasons are public reasons, and so perhaps she has moved away from the claim that the publicity of reasons is akin to the publicity of language. Perhaps in *Self-Constitution* a public reason *is* an agent-neutral reason.[12] Let's look at what else she has to say about public reasons in *Self-Constitution*.

In the immediate paragraph after the passage in which Korsgaard identifies a public reason as both an agent-neutral reason and a reason "whose normative force extend[s] across the boundaries between people", she writes the following:

> As many philosophers have pointed out, the privacy of reasons is consistent with a kind of universalizability requirement. If I conceive of reasons as private, and accept a universalizability requirement, I am committed to the view that if I have a reason to do action-A in circumstances C, then I must be able to grant that you also would have a reason to do action-A were you in circumstances C. So for instance, if I think that the fact that something will make me happy is a good reason for *me* to do it, then universalizability requires me to think that the fact that something will make you happy is a good reason for *you* to do it. But my happiness is still mine, and yours is still yours; mine is a source of reasons for me, but not for you; yours is a source of reasons for you, but not for me. On the public conception of reasons, by contrast, a universalizability requirement commits me to the view that if I have a reason to do action-A in circumstances-C, I must be able to *will* that you should do action-A in circumstances-C, because your reasons are normative for me. (2009, p. 191)

[12] Remember: the significance of this concerns what Korsgaard thinks is required for performing a collective action. If public reasons are agent-neutral reasons, then two people can share their agency only if they act on agent-neutral reasons.

This is a very confusing passage. Before we analyze it piece by piece, we should note that its aim seems to be to further characterize private and public reasons in terms of the different ways in which they satisfy a universalization constraint.

Let's take the first sentence of this passage:

(1) As many philosophers have pointed out, the privacy of reasons is consistent with a kind of universalizability requirement.

It's not clear who Korsgaard has in mind when she says "many philosophers". She doesn't cite anyone here. Perhaps she has someone like Joshua Gert in mind, who we just saw has argued that the publicity constraint is satisfied by agent-relative reasons no less than agent-neutral reasons. But what does publicity have to do with universality?

We saw in the previous chapter that Korsgaard holds principles to be universal by their very nature, for to be a principle is to be applicable in a wide variety of situations. We also saw that when Korsgaard talks about adopting a principle for action, this is another way of talking about acting for a reason. If principles, by their very nature are universal, then so are reasons. But applicability across a wide variety of situations is another way of articulating the publicity constraint which Wittgenstein says is constitutive of language and which Korsgaard says (in *Sources*) is constitutive of reasons. So when Korsgaard suggests that the privacy of reasons is consistent with a kind of universalizability requirement, it would appear that she is conceding that private reasons are public in Wittgenstein's sense. So "public" as she is using the term in *Self-Constitution* must be different from how she is using it in *Sources*.

Let's now turn to the second sentence of this passage:

(2) If I conceive of reasons as private, and accept a universalizability requirement, I am committed to the view that if I have a reason to do action-A in circumstances C, then I must be able to grant that you also would have a reason to do action-A were you in circumstances C

There is something odd about the way this is phrased. The universalizability requirement is couched in terms of how the individual *conceives* of her reasons and whether the individual *accepts* the universalizability requirement. This raises the possibility that one need not conceive of their reasons as private or public and that one need not accept the universalizability requirement on their reasons. But we know from her argument against particularistic willing that Korsgaard holds that we cannot have "one-off" reasons, that is, reasons that apply only in a particular situation and in principle never again. It is incoherent to not accept a universalizability constraint.

Perhaps when Korsgaard is talking about how one conceives of reasons and whether one accepts a universalizability constraint, she is not talking about the proverbial practical-reasoner-on-the-street but rather philosophers who might disagree about the nature of practical reasons. So when she says, "if I conceive of reasons as private", who she really has in mind is someone like the Hobbesian who thinks that practical reasoning is essentially an economic endeavor in which one's aim is to maximize expected utility; and when she says "and accept a universalizability constraint" she again has in mind a philosopher theorizing about the nature of reasons.

If we read Korsgaard in this way – as talking about philosophers theorizing about reasons – the sentence becomes less opaque. Consider Korsgaard's next sentence, which I take it is supposed to be an example of a quintessential private reason:[13]

> (3) So for instance, if I think that the fact that something will make me happy is a good reason for *me* to do it, then universalizability requires me to think that the fact that something will make you happy is a good reason for *you* to do it.

Let's make this more concrete. As I'm writing this I am looking out the window. It is a beautiful day in Riverside. On beautiful days, leisurely walks make me happy. The fact that it is a beautiful day, combined with the fact that leisurely walks on beautiful days make me happy, together provide me with a reason to go take a walk. If we universalize this, then we find that insofar as talking leisurely walks makes *you* happy, then you too have a reason to take a walk today. There is a kind of equation here for private reasons: similar inclinations + similar situations = similar reasons.

But what makes these reasons *private*? Notice that in the example provided we find an agent-relative reason: this makes *me* happy. It is not the happiness which is important, so much as it is *my* happiness. There is an essential reference to the agent whose reason it is. This, I take it is what Korsgaard is highlighting in the next sentence:

> (4) But my happiness is still mine, and yours is still yours; mine is a source of reasons for me, but not for you; yours is a source of reasons for you, but not for me.

It would seem, at least at this point, that private reasons are indeed agent-relative reasons. But are public reasons agent-neutral?

---

[13] We still don't know what a private reason is, exactly, except, perhaps, an agent-relative reason or a reason whose normative force is not shared, whatever that might happen to mean.

Let's look now at the last sentence of this passage:

(5) On the public conception of reasons, by contrast, a universalizability requirement commits me to the view that if I have a reason to do action-A in circumstances-C, I must be able to *will* that you should do action-A in circumstances-C, because your reasons are normative for me.

Things are not as clear in this passage as they were in the previous two passages. Recall that Korsgaard states that a public reason is a reason "whose normative force extends across the boundaries between people". Up to this point, we have been focusing on the question of whether public reasons are nothing but an agent-neutral reason. But perhaps what is more crucial to the idea of a public reason is the obscure idea of a reason's normative force being shared. Might passage (5) be an attempt by Korsgaard to unpack what she means by this?

What might it mean to say that a reason's normative force extends across the boundaries between people? Presumably it is to say that if you have a reason to do something, you are not the only one who feels the normative force of this reason. Perhaps I do as well. Agent-neutral reasons certainly have this feature. Since we all have a reason to reduce suffering when doing so is not overly taxing, it follows that if you have such a reason, then I feel the normative force of the reason too (since I as well have the reason).

But how does this idea connect to the universalization constraint stated in (5)? Korsgaard writes that the constraint commits one to holding that if one has a reason to A in C, then one must *will* that others should do A in C as well. And she says that this is a consequence of recognizing another person's reasons as normative. But why is this?

To answer this question we need to contrast the universalization constraint as applied to public reasons with the universalization constraint applied to private reasons.

When applied to private reasons, we find that if S has a reason to ϕ in circumstances C then S is committed to recognizing that anyone who finds themselves in C also has a reason to ϕ. When the constraint is applied to public reasons, we find that if S has a reason to ϕ in circumstances C, then S is committed to *willing* that anyone in C also has a reason to ϕ. How is S's *willing* that others share her reason different from S *recognizing* that others share her reason? What difference does *willing* make? Perhaps reflection on the following passage will help:

> So on the private conception of reasons, a universalizability requirement leaves us each with our own system of private reasons, which don't have to be consistent with anyone else's. And this can leave us in a condition of essential conflict. For instance, suppose you and I are competing for some object we both want. I think I have a reason to shoot you, so that I can get the object. On the private conception of reasons, universalizability commits me to thinking you also have a reason to shoot me, so that you can get the object. I simply acknowledge that fact, and conclude that the two of us are at war. Since I think you really do have a reason to shoot me, I think I'd better try very hard to shoot you first.
>
> But on the public conception of reasons, we do not get this result. On the public conception I must take your reasons for my own. So if I am to think I have a reason to shoot you, I must be able to will that you should shoot me. Since presumably I can't will that, I can't think I have a reason to shoot you. So it is only on the public conception of reasons that a universalizability requirement is going to get us into moral territory. (SC 191-192)

It might at first seem that in this passage Korsgaard is suggesting, unlike with private reasons, that the normative force of public reasons extends beyond the borders of persons. In her example, person A and B are after the same object, only one of whom can have it. One way for A to get the object is for A to kill B. Since B is in the same circumstances as A, he must have reason to kill A. This puts A and B at a standstill. Because the reason, so conceived, is private, the normative force of A's reason is not felt by B. If it were, B would have a reason to promote A's shooting of him. This might be

what Korsgaard means when she says that private reasons are not shared: their normative force is felt only by the individual whose reason it is. On this reading, a public reason is a reason whose normative force is felt by everyone. If A's reason to ϕ is public, then everyone has a reason to promote A's ϕ-ing.

But this can't be what Korsgaard means by public reason. When she imagines A's reason to shoot B is a public reason, she states, not that B has a reason to promote B's being shot, but rather that "[A] must be able to *will* that [B] should shoot [A]." Korsgaard then suggests that this would introduce a kind of contradiction in A's will: "Since presumably I can't will that, I can't think I have a reason to shoot you." Why can't A will this? Presumably, A is willing that she get the object that both A and B are after. If she wills that B should shoot her, then she is effectively willing both that she should get the object and that she should not.

This idea is reminiscent of the reading of Kant advanced by Rawls (2000), Hermann (1996), and Korsgaard (1996b), among others, that championed what has come to be called the "contradiction in the will test". This "test" was an interpretation of passages of the *Groundwork* in which Kant discusses maxims that cannot be universalized (e.g., deceitful promises and maxims of indifference). According to this interpretation, a maxim is deemed permissible to act on only if it can consistently be willed as a universal law without undermining the end one is trying to achieve. Perhaps Korsgaard's idea is that public reasons are those that pass this test.

If this is what Korsgaard means by public reason, it is unclear what the connection is to agent-neutral reasons, at least as they are defined by Parfit and Nagel. Do

only agent-neutral reasons pass the test of universalization? The answer to this question is certainly no. It would be implausible if *no* agent-relative reasons passed the CI-test.

At this point I'd like to reconsider the perspectival aspect of the passage. Earlier I suggested that we treat the perspectival language as appealing to a *theorist's* perspective. But that seems to not have gotten us very far. It is significant that unlike with most discussion of agent-neutral and agent-relative reasons, the reasons under considerations are not described from a God's eye point-of-view. In the passages, Korsgaard *does* seem to ask us to consider the perspective of an individual's practical reasoning. She writes, "If I conceive of reasons as private … then I am committed to the view that you also would have a reason." She goes on to add, "On the public conception I must take your reasons for my own." I conjecture that the key to understanding Korsgaard's idea of public reasons might lie in the idea that they describe, not an ontological feature that some reasons have while others lack, but rather a *method* of practical reasoning. A public reason is a reason approached in a certain way in practical reasoning. To treat a reason as public, is to reason in public. To treat a reason as private is to reason in private.

What is the difference between reasoning in public and reasoning in private? To reason in private is to determine only whether your reasoning is internally consistent independently of others. The ends of others do not play a role in private reasoning. But reasons so understood still have a universal aspect to them – and they must, since reasons by their very nature are principles or rules that describe what to do given certain circumstances. If I am a private reasoner, I can recognize your reasons as reasons, but

they just do not play a role in my practical reasoning (other than, perhaps, as road blocks, or things that get in my way).

Reasoning in public on the other hand, involves more than just determining whether your reasoning is internally consistent. When you reason in public, you consider the ends of others. Does my reason bring me into conflict with the ends of others? If I give my reason to others, can I still achieve my end? If I can't, if my reason requires making an exception out of myself, then I am acting on reasons that others will not consent to. And if others will not consent to these reasons – if they would reasonably reject my reasons – then I am not reasoning in public.

I believe that this way of understanding the private/public distinction helps us also to also make sense of the connection that Korsgaard wants to draw between public reasons and what she calls shared deliberation. Public reasons, so understood, are the constituents of public reason*ing*. Public reasoning, I have been saying is practical reasoning one engages in with others. And this, I take it, is what shared deliberation is.

So public reasons are the constituents of public reasoning. What does this have to do with agent-neutral reasons? Presumably, all agent-neutral reasons are going to be public reasons: they will be reasons that can fit into shared deliberation. But are all public reasons agent-neutral reasons? I think the answer to this question is no. Consider the following case.

> **Neighborhood Garden** Gary and Joanne, who are members of a neighborhood cooperative garden, are trying to figure out what to plant in an open spot in the garden. In making this decision they've listed a number of considerations that they must take into account: the likelihood that the plant will be able to grow in the community's garden, whether their community will enjoy the crops, and the amount of labor that the community will have to invest in helping the plant grow.

> They decide to opt for carrots as they think that this will maximally satisfy these three desiderata.

Gary and Joanne engage in shared deliberation in a joint effort. In doing so, they reason publicly in that they try to make their ends consistent with those of the community garden. Are the reasons that they consider agent-neutral? Consider the reason associated with the fact that the community will enjoy the carrots. Recall that agent-neutral reasons are ones that do not make essential reference to the point of view of the individual whose reason it is. In this case, the reason is associated with the point of view of the community. It is a reason for the *community* to plant carrots – which is why it factors into Gary and Joanne's practical reasoning – but it is not a reason for me or you to plant carrots, since we are not members of this community.

This community-relative reason is not an agent-neutral reason. But is it a public or private reason in Korsgaard's sense? Consider again the two different glosses Korsgaard gives us for what a public reason is:

> Our reasons must be what I call public reasons, reasons whose normative force can extend across the boundaries between people. Public reasons are roughly the same as what are sometimes called objective, or agent-netural reasons. (2009, 191)

The community-relative reason we just considered is not an agent-netural reason. But it is a reason "whose normative force can extend across the boundaries between people". It is a reason that is shared. It is hopefully now evident that while all agent-neutral reasons might be reasons that are shared, not all shared reasons are agent-neutral reasons.

But we still haven't settled whether we should consider the community relative reason as a public reason or not. The answer to this question is that the concept of a

public reason is not well-enough defined to say. It depends which of the two notions Korsgaard intends 'public reason' to pick out. This is problematic for Korsgaard's *Collective Action Argument*.

3. Problems with the Collective Action Argument

The question that we are currently interrogating is how Korsgaard purports to establish that the moral law is constitutive of shared agency. In section 1 we saw that she offers the following argument, which I have been calling the *Collective Action Argument*:

1. Collective action requires shared deliberation.
2. Shared deliberation requires acting on public reason.
3. Acting on public reason requires complying with the moral law.
4. So, collective action requires complying with the moral law.

In section 2 we saw that the concept of public reason is not well-defined by Korsgaard. In fact, she offers two distinct – overlapping, but not coextensive – glosses on what she means by a public reason. On one hand they are reasons that are shared – they are the kind of thing that factor into shared deliberation; on the other hand, they are agent-netural reasons. But the latter, we saw, is a *proper subset* of the former. This presents a problem for the Collective Action Argument. In premise 2, the concept of public reason being used is that of a reason that is shared. In premise 3, the concept of public reason is that of an agent-netural reason. So, the argument that Korsgaard is *really* offering is the following:

1. Collective action requires shared deliberation
2. Shared deliberation requires acting on reasons that are shared.
3. Acting on agent-netural reason requires complying with the moral law.
4. So, collective action requires complying with the moral law.

Though premise 2 and 3 are likely true, the argument so understood is obviously invalid. Korsgaard's attempt to show that the moral law is a constitutive principle of shared agency fails.

Nevertheless, I think there is there another way to defend the Shared Agency Hypothesis. My aim in Part II of this chapter will be to do just that.

Part II: In Defense of the Shared Agency Hypothesis

In Part I of this chapter I argued that Korsgaard's *Collective Action Argument* is invalid. Insofar as this is the crucial argument for securing the foundations of morality, this throws her entire project into doubt. My aim in the rest of this chapter is to defend the Shared Agency Hypothesis by offering a different kind of argument in its defense, one that doesn't rely on the concept of public reason.

To defend the Shared Agency Hypothesis I will investigate the metaphysics of shared agency.[14] We will discover that shared agency is a kind of normative relationship. I will argue that entering into and maintaining this normative relationship requires treating the other parties in the relationship as ends in themselves. Insofar as this is the content of the fundamental principle of morality, it will follow that the moral law is a constitutive principle of shared agency, in the sense of "constitutive principle" that I set

[14] I will be using "shared agency" and "collective action" interchangeably depending on which fits better in the sentence stylistically. Later in the chapter I will by using shared intention as a proxy for shared agency. The reason I do this is because much of the literature on the issue concerns the question of what a shared intention is. I will be assuming that the conditions for sharing an intention are roughly the same as the conditions for sharing agency, the later requiring in some sense the execution of the former.

out in chapter 1: a principle which describes and governs the activity we call collective action.

Before getting underway, I'd like to highlight a few obvious worries that will no doubt pop into the head of most people. The most obvious worry is that my strategy will not be able to capture the categoricity of morality. This strategy, the worry goes, will force me onto the second horn of Katsafanas' Dilemma discussed in the previous chapter. This problem is one worth worrying about. I will consider it in the concluding chapter.

The second worry I'd like to highlight is that not treating others as a mere means is only *part* of morality. If this is all my argument secures, then it can hardly be said that I've secured a foundation for *morality*. I think this is right. There is more to morality than just this. Perhaps we have obligations to animals and, depending on our theory of moral status, duties to the environment. We might also think that morality involves more than obligations. Perhaps there are ways of being good that are not obligatory.[15] Maybe so, but I think not treating each other as mere means is the *core* of morality, and I'm inclined to think that it is the most important part of it.[16] If the biggest criticism my view faces is that I've only provided a foundation for the core of morality, I will consider it a success.

Without further ado, let's get started.

---

[15] Sometimes this category of morality is called "the supererogatory".

[16] Following Scanlon (1998), we might say that what I am seeking to provide a foundation for is "morality in the narrow sense" or "what we owe to each other". This I am happy to concede.

4. A Sketch of the Argument

Before jumping straight into the details of the arguments, I would like in this section to provide a sketch of the argument I will be providing in favor of the Shared Agency Hypothesis in order to elicit some intuitions in favor of it. What the Shared Agency Hypothesis proposes is that, in order to perform even mundane collective acts – such as dancing the tango or loading a van – the parties to the action must not treat each other as mere means. There has been quite a lot written on the phenomenon of shared agency, and very few have defended something approaching the Shared Agency Hypothesis.[17] Though radical, I also think the Shared Agency Hypothesis has a lot of intuitive support. My aim in this section is to provide this intuitive support.

4.1 Cases

Let's begin by reflecting on some cases.

> **Happenstance** Kaleigh and Katie don't know each other, but have a mutual friend, Juan, in common. By happenstance, both Kaleigh and Katie decide to surprise Juan by painting his house on Saturday. (Juan has been complaining about his house's paint for a while.) Kaleigh and Katie never encounter each other and each only paint one side of the house before giving up (because it was hot).

Unbeknownst to each other, Katie and Kaleigh are both painting Juan's house. Intuitively, however, Katie and Kaleigh are *not* painting Juan's house *together* in the sense that Katie and Kaleigh are sharing their agency and engaging in a collective act. They are acting individually with a common aim.

[17] As far as I can tell, Korsgaard is the only such philosopher.

Now consider the following case:

> **Deceptive Painting** Chris wants Sara to help him paint his house. But he knows that Sara – who doesn't like to leave jobs unfinished, but also has pressing matters later in the day – won't do it unless she is deceived about how long the painting will take. Chris tells Sara it will only take an hour, when in fact it will take three. Because Sara thinks the house will only take an hour to paint, she joins Chris to paint the house.

Are Chris and Sara engaged in a collective act of painting the house? Are they *sharing* their agency? It might, at first glance, appear that they are. After all, Chris and Sara both *intend* to paint the house. In all outward appearances they seem to be painting the house *together*. And if you were to ask each of Chris and Sara what they are doing, they would both likely exclaim that they are doing so. But I think this is a mistake. We are lulled into saying this by outward appearances. Chris is using Sara in a way not too different from the way he is using his paint roller. His paint roller helps him paint more of the wall than a paint brush would. Sara is doing the same for him. Sara is not *sharing* her agency with Chris. Rather, she is being *used* by Chris. She is an *instrument* for Chris in the pursuit of his end.

Of course, Sara *is* using her agency to paint the house. She intends to paint the walls of the house and takes the means to this end. But compare the case of **Deceptive Painting** with that of another, which I will call **Painting Together**:

> **Painting Together** Chris wants Sara to help him paint his house. But he knows that Sara – who doesn't like to leave jobs unfinished, but also has pressing matters later in the day – won't do it if it will require three hours of her time. Knowing that Sara only has an hour to spare that day, Chris asks Sara if she would be willing to show up for the last hour that he will be painting the house. Sara agrees to this and spends an hour of her day painting the house with Chris.

Outwardly, Sara and Chris will look like they are engaged in the same activity in **Painting Together** as they are in **Deceptive Painting**. But I will argue that they are not. In **Painting Together**, Sara and Chris are genuinely working together – they are sharing their agency in a joint effort. In **Deceptive Painting**, by contrast, Sara's agency has been *appropriated* by Chris. In treating Sara the way that he does, Chris takes away Sara's ability to *consent* to the activity. She is being deceived about the true nature of the effort, and in so being deceived, she has a false conception of what she is doing.

Maybe your intuitions aren't with me yet. Consider a far more extreme example:

> **Slavery** Alfred is an overseer on a plantation. He orders a slave, Jim, to help him paint a house. Jim, knowing that Alfred will beat him if he doesn't help, joins Alfred in painting the house.

Are Alfred and Jim painting the house together? It is beyond arguable that Alfred is using Jim, even more so than Chris is using Sara. Alfred is using force on Jim to get him to contribute labor to the painting of the house. In so forcing Jim, Alfred nullifies Jim's agency appropriating it to his own end. Of course, like Sara, Jim is not a robot. He is performing genuine actions in performing the labor. However, he and Alfred are not painting the house *together* in the sense that their agency is *shared*. Alfred has *taken over* Jim's agency. Jim has become an instrument of Alfred's.

In the cases of **Deceptive Painting** and **Slavery** we find individuals who have had their agency appropriated by others. These sorts of cases are not ones in which the parties are *sharing* their agency with each other. While it's true that in both of these cases the individuals are both painting the house, they are not painting the house *together* in the sense that they share their agency.

Cases that look in all outward appearances like a case of shared agency but turn out not to be ones when we look "under the hood" are not all that uncommon. **Happenstance** was one such case. What I will be arguing in this chapter is that cases where one person uses another *as a mere means* in the pursuit of an end are cases that may look in all outward appearances like a case of shared agency, but when we look "under the hood" we found out that they are not. The reason that these turn out not to be cases of shared agency is that using another as a mere means involves an appropriation of that person's agency such that this second person's agency becomes a mere extension of the first person's. Shared agency, by contrast, requires the autonomous exercise of the agency of all parties. This is required in order for the relational norms constitutive of shared agency to be instantiated.

In order to make my case that this is the correct picture of shared agency, I need to explain how I will be understanding the phrase "treat as a mere means". This will be my next task. Following that I will dive into the two main players in the literature on shared agency, Margaret Gilbert and Michael Bratman, and show that, after reflecting on their respective theories, we ought to accept the Shared Agency Hypothesis.

5. Treating as a mere means

Kant's so-called Formula of Humanity tells us to treat others always as an end, never merely as a means. To those not trained in moral philosophy, this formula can seem a bit opaque. If we are going to treat the formula of humanity as our fundamental moral principle – and especially if we are going to say that it is a constitutive principle of a

certain kind of activity – we are going to have be clear about what it actually says. Unpacking this principle will be my aim in this section.

Generally it is recognized that the term "means" in the phrase "mere means" refers to a means to an end. To say that we shouldn't treat others as mere means, then, is to say that we shouldn't treat others as mere means *to an end*. A common example of something that we use as a means is a tool or instrument such as a hammer. A hammer is something we use *for* hammering nails. When our end is to secure two planks of wood together with a nail, we use a hammer to achieve this end. A hammer – in being nothing more than an instrument – is a *mere* means.

As with a hammer, we treat a person as a means when we do something to them in order to achieve an end. When I ask my roommate for a ride to the airport, or to pick up dish washer detergent at the grocery store, or to return a book at the library for me, I treat him as a means. (What I "do" to him is ask a request of him.) But my relationship with my roommate is such that I am not acting *wrongly* when I do so. It is not always wrong to treat someone as a means. According to the Formula of Humanity, treating a person as a means is wrong when we treat them *merely* as a means, when we treat them as we would a hammer. But what is it to treat them in this way?

In the voluminous literature on this question there are two major views about what is involved in treating another as a mere means: consent views and end-sharing views. For the purposes of this dissertation, I will be borrowing Samuel J. Kerstein's (2013) Hybrid Account which attempts to synthesize these two views, thereby combining all of their virtues and none of their vices. I won't be offering a full defense of Kerstein's

Hybrid Account; the literature on this issue is so immense that it would require its own dissertation. In place of a full defense of the Hybrid Account I will offer a truncated version of Kerstein's case-based argument for it.

5.1 Consent Views

The most popular account of what it is to treat another as a mere means is what has become known as consent accounts. There are two varieties of consent accounts: *actual* consent accounts and *possible* consent accounts. According to *actual* consent accounts, individual A treats individual B as a mere means if A treats B in a way that B does not consent to. This kind of view is called an actual consent account because it requires that we secure *actual consent* from individuals in order to treat them as we wish to. Some philosophers, identifying problems associated with determining when genuine consent is provided, argue that, rather than looking to secure consent from those we interact with, we should instead focus on providing people with a genuine opportunity to *dissent* from our treatment of them. This kind of view is called a *possible consent* view because it holds that we treat others as mere means when we treat them in a way that they cannot possibly consent to; and people cannot consent to being treated in a certain way when they lack the opportunity to dissent from that treatment.

Perhaps the most prominent philosopher who adopts the actual consent view is Robert Nozick, who provides a brief articulation of it in *Anarchy, State and Utopia*:

> Side constraints upon action reflect the underlying Kantian principle that individuals are ends and not merely means; they may not be sacrificed or used for the achieving of other ends without their consent. (1974, pp. 30-31)

The actual consent account does not conflict with common sense. In fact, it is perhaps a truism of popular morality that we should treat people only in ways that they consent to.

Actual consent views, however, have been criticized by Kantians, most famously by Onora O'Neill. In her classic paper, "Between Consenting Adults", O'Neill argues that actual consent views are problematic because they paint over several of the difficulties associated with securing consent. She points out that in many contexts it is often unclear what consent amounts to. Is signature or verbal agreement sufficient? How about a nod of the head? O'Neill points out that we can secure the outward appearances of consent – or what she calls "the formal indicators of 'consent'" (p. 107, fn 1) – while still treating those we interact with as mere means. For example, we can coerce people into signing a contract or press someone in a vulnerable position to accept a raw deal. Focusing on the outward manifestations of consent may lead us to accept "superious consent" as actual consent (p. 107).[18]

A second problem associated with focusing on actual consent is that what individuals seemingly consent to sometimes does not match the behavior they are subsequently subjected to. For example, I might agree to a certain job for a certain wage because I don't fully understand the inherent risk involved in the labor I will be asked to perform. (Perhaps I don't understand that if I work in the coal mine my chances of getting black lung go up significantly.) However, in securing my signature on the

[18] This problem that O'Neil points to is really an epistemic problem and not a problem associated with actual consent *per se*. Presumably, each of the examples just mentioned would not count as examples of consent secured in the actual consent account.

contract, you might think that you've got all you need to satisfy what morality demands of you on the labor market.

The final problem that O'Neill points out for actual consent accounts is that they fail to adequately take into consideration how impairment can affect one's ability to offer genuine consent. When under duress or pain, I might consent to treatment that I wouldn't in normal circumstances.

In criticizing actual consent views, O'Neill is not suggesting that consent has no moral significance. For O'Neill, what is morally significant about consent is not the word of approval *per se* that we get from those we interact with, but rather how we *approach* those we interact with. When we interact with others, we ought to approach them as agents with the capacity to set their own ends through practical reason. O'Neill suggests that when we interact with persons what we ought to concern ourselves with is providing them with the space to dissent from our treatment. For example, when we have pressed someone in a poor bargaining position *because* they were in such a position what we have done is identified an individual who lacks the proper normative space (so to speak) to dissent, *and for this reason* we have decided to so interact with them.[19] By taking advantage of the fact that the person is in a poor bargaining situation, we are, in effect, not allowing them the possibility to dissent. This means they lack the proper normative space to fully exercise their agency. By approaching them in this way we have thus

[19] Here and elsewhere when I speak of "normative space", I am referring to the genuine options one sees available to one in practical reasoning. Normative space, so understood, is perspectival and subjective. When I refer to an individual's normative space, I generally have in mind the normative space that a "reasonable person" would see themselves in within the context under discussion. The reasonable person I have in mind is the same "reasonable person" that the so-called reasonable person test in the law has in mind.

neutralized their agency, so to speak, and in neutralizing their agency, we treat them like an instrument or tool in the pursuit of our end. We know they are in a poor bargaining situation, so we know that they lack the ability to dissent from, e.g., signing a contract to be coal miners in our unsafe coal mines. In treating our prospective employees in this way, we treat them as mere instruments for the extraction of coal from our mines, not much different from the pick axes we force them to buy from the company store (which they must do so to operate in the company mines).

Recognizing that the possible consent account is *modal*, one might think it has the unwelcome implication that we can treat others in ways that they don't consent to so long as they *could* consent to it. This is a mistake. If one treats someone in a way that they don't consent to, then one *ipso facto* treats them in a way that they *can't* consent to. Consider the case where you *do* in fact dissent to the way I wish to treat you, but I treat you accordingly anyhow. One might think that since you *did* dissent, you had the chance to, and so were given such a possibility. The problem here is that if your dissent has no purchase on my practical reasoning, then I in fact give you *no* genuine opportunity to consent. In order to give you such an opportunity, I must allow your dissent to alter my course of action. Possible consent accounts, then, should not be construed as placing no importance on consent.

5.2 End Sharing Views

Before considering some of the problems facing the possible consent account, I'd like to introduce the second of the two popular accounts of what is involved in not treating

someone as a mere means: the end-sharing account. This way of understanding what is involved in treating someone merely as a means is inspired by Kant's discussion of false promising in the *Groundwork*:

> As far as necessary or owed duty to others is concerned, someone who has it in mind to make a lying promise to others will see at once that he wants to make use of another human being *merely as a means*, [the latter of which] does not at the same time contain in himself the end. For the one I want to use for my purposes by such a promise cannot possibly agree to my way of proceeding with him and thus [cannot] himself contain the end of this action. (1785/2012, pp. 429-30)

Kant's suggestion here seems to be that one uses another as a mere means if one uses another for an end that the latter "cannot himself contain". But what does it mean for one to be unable to contain an end?[20] Kerstein suggests the following as a first pass at unpacking what might be going on here:

> **End-sharing Pass 1:** if another cannot share an agent's end in using her in some way, then the agent treats the other merely as a means. (Kerstein, p. 60)

This first pass at the end-sharing account raises two questions. First, what does it mean for two individuals to share an end? Second, what does it mean to say that two or more individuals *cannot* share an end? According to Kerstein two individuals share an end if "they are both trying, or at least have both chosen to try, to realize this end" (ibid). The second question is a trickier one for Kerstein. In an effort to understand it, Kerstein runs through six possible interpretations, offering counter-examples to each. I will spare the reader and not run through each of these. Instead I will consider what I take to be the

[20] It should be noticed that in this passage we actually find the two interpretations we are entertaining of what it means to treat merely as a means. The phrase "cannot possibly agree" can easily be interpreted as indicating the possible consent account discussed in section 5.1.

three most intuitive interpretations. Considering these interpretations will help us to better understand the Hybrid Account and why it is worth endorsing.

The most straightforward way to understand what it means to say that two individuals cannot share an end is perhaps the following:

> **Logical Impossibility Account:** P cannot share S's end if and only if it is logically impossible for P to share the end that S is pursuing**.**

Kerstein rightly points out that this account cannot capture the false promising case that prompts it. Imagine that I borrow money from my mother promising to repay it, but with no intention of doing so. It is certainly not *logically impossible* for my mother to have the end of providing me with money that I do not need to repay. In fact, my mom might have the intention of refusing to allow me to repay the money were I ever to try to do so. Though this is a clear case of using another as a mere means, the Logical Impossibility Account cannot capture it.

If 'cannot' in the phrase 'cannot share an end' does not refer to logical impossibility, to what could it refer? Kerstein suggests that we could interpret it to mean "practically irrational". To say that P cannot share S's end, then, might perhaps mean the following:

> **Practical Irrationality Account** P cannot share S's end if it would be *practically irrational* for P to adopt S's end where 'practically irrational' refers to instrumental irrationality (i.e., means-end incoherence).

What does the practical irrationality account have going for it? Consider yet again our false promising case, but this time with a bit more detail added:

> **Cell Phone** Pam has been saving up her money to buy a new cell phone. Her friend Saul also wants a cell phone but has no money. Saul asks to borrow money

> from Pam so that he can buy some drugs, however he has absolutely no intention of paying a cent of it back (he doesn't care he if burns that bridge).

In **Cell Phone,** Saul's end is to buy drugs for himself with Pam's money and never repay her. It would be *irrational* for Pam to adopt Saul's end for doing so would make her means-end incoherent. The practical irrationality account provides us with a very concrete way of spelling out what it means to be unable to share an end.

However, the Practical Irrationality Account is not without problems of its own. Imagine that Saul intends to buy some marijuana to help with his glaucoma. Perhaps Pam, were she to understand this, would decide that his ends are more important and that hers are worth giving up. (Perhaps she would think, *he needs that marijuana more than I need a new cell phone*.) Were she to do this – were she to give up her end upon hearing why Saul "borrowed" the money – she would no longer be irrational in adopting Saul's end, for she would no longer be means-end incoherent. But if Pam can adopt Saul's end without being means-end incoherent, then it would turn out, according to the Practical Irrationality Account, that Saul is *not* using Pam as a mere means. Should Saul be able to get off the moral hook so easily?[21] It seems not. This line of thought prompts Kerstein to further complicate the practical irrationality account as so:

> **Refined Practical Irrationality Account:** P cannot share S's end E if and only if P has an end F such that P's pursuing F at the same times as E would (1) violate the instrumental principle, and (2) P would be unwilling to give up pursuing F even if (a) P was aware of the likely effects of S's pursuit of E, and (b) P was

[21] One could even imagine Saul reasoning like so: were Pam to find out that I was going to buy a new phone with the money, she would share my end. But I'm going to lie about it anyway, because I don't want her to feel implicated in an illegal action, and I know that she would prefer not to know what I'm doing with the money. If Saul reasons like this, we might think that he is *not* using Pam as a mere means, even though he is lying to her. Saul isn't using her as a mere means because Saul *knows* that Pam shares his end.

> aware that S would give up pursuit of E based solely on P's preference that S not pursue it.[22]

This is very complicated so it deserves elaboration. Perhaps the best way to unpack the refined practical irrationality account is to use the case that motivates Kerstein to develop:

> **Mugging**: A loiterer threatens a passerby with a gun, demanding that the passerby hand over their wallet. The passerby is on her way to a movie. If she hands over her wallet she will be unable to attend the movie.

In **Mugging,** the loiterer is using the passerby as a mere means. The passerby would, of course, violate the instrumental principle were she to adopt the mugger's end of handing over her wallet, for doing so would conflict with her end of attending the movie. But the passerby could abandon her end. She might recognize (correctly) that if she is to attend the movie tonight she will have to somehow resist the mugging. Since she doesn't want to do that, she might give up her end of going to the movie. But if she does this, it will no longer be practically irrational for her to adopt the muggers end.

What we need to fix the practical irrationality account is a condition that can somehow keep the victim's ends prior to the interaction fixed. This is what condition (2) in the **Refined Practical Irrationality Account** does:

> P would be unwilling to give up pursuing F even if (a) P was aware of the likely effects of S's pursuit of E, and (b) P was aware that S would give up pursuit of E based solely on P's preference that S not pursue it.

In **Mugging** the passerby would be unwilling to give up her end of going to the movie even if (a) she becomes aware that the mugger will harm her as he pursues his end of

[22] This is not a direct quote of Kerstein. This way of articulating the account is, I hope, much clear than his.

getting her wallet, and (b) she becomes aware (contrary to fact) that the mugger would give up the pursuit of his end based solely on the passerby's preference that he do so. Since the passerby would prefer that she both not get mugged and go to the movies, both conditions (a) and (b) are satisfied.

Let's return to **Cell Phone**. Can the **Refined Practical Irrationality Account** deal with **Cell Phone** where the unrefined account failed? It would seem not. We saw that Pam would be willing to give up pursuing the purchase of a cell phone even if she was aware of the likely effects of Saul's pursuit of his purchasing pot (namely that she would be lied to about the purchase) and if Saul would give up his end if Pam so wished (because, again, were Pam to know what Saul is putting his money to, she would think that this end is more important than hers). To solve the problem still plaguing the practical irrationality account, we need to add one more refinement:

> **End Sharing Account (Final Pass):** An agent treats another merely as a means if it would be *unreasonable* for the agent to believe that the other can share the proximate end or ends the agent is pursuing in treating him as a means. (The notion of end sharing is that which is captured in the **Revised Practical Irrationality Account**.)

According to **End Sharing Account (Final Pass)** Saul treats Pam as a mere means if it is *unreasonable* for him to believe that Pam can share his end (as this is understood in the **Revised Practical Irrationality Account)**. Presumably it would not be reasonable for him to believe that Pam would give up her end and embrace his.[23]

[23] What if it *was* reasonable for him to believe this? Perhaps Saul knows Pam really well, and knows that she would forgo a new cell phone in order for him to get pot for his glaucoma. But perhaps Saul also knows that Pam would prefer to believe that she will be getting the money back, for it will make it easier for her to part with the money, and so she would prefer that he lie about his intent to pay her back. In this sort of case, it's not obvious that Saul *is* using Pam as a mere means. Quite the contrary, Saul is being quite considerate about Pam's feelings.

5.3 Hybrid Account

Both the **Possible Consent Account** and the **End Sharing Account** face counterexamples. Let's start with the counterexample to the End-Sharing Account:

> **Competitive Tennis**: Pete and Andre are competing in the men's singles final at a season-ending tennis tournament. Both players are going to retire after the tournament, and each player has as his goal to end his career by defeating the other. (Kerstein, p. 66)

In **Competitive Tennis** Pete has the end of defeating Andre and winning the tournament. Andre has the end of defeating Pete and winning the tournament. Were Pete to adopt Andre's end, Pete would be instrumentally irrational, for his new end would conflict with the end he currently has of winning the tournament. Similarly, were Andre to adopt Pete's end, Andre would be instrumentally irrational for the same reason. So, neither one can share the other's end. However, as Kerstein points out, it seems wrong to say that Pete and Andre treat each other as mere means.

Consider now the following counterexample to the Possible Consent Account:

> **Surprise Party**: You are throwing a surprise party for your partner, who loves surprise parties. Your sister in law knows that your partner loves surprise parties but is also terrible at keeping secrets. Knowing this about your sister-in-law you lie to her about the party plans.[24]

In **Surprise Party** you have rendered your sister in-law unable to consent to your treatment. According to the Possible Consent Account, this means that you are using her as a mere means. But again, this seems wrong.

Though the Possible Consent Account gets **Surprise Party** wrong, it gets **Competitive Tennis** right. Pete and Andre may not be able to share each other's ends,

[24] This is a case that comes from Kerstein, but I've spelled it out in a different way than he has.

but they do each consent to the other's treatment. (They consent to abiding by the rules of tennis and consent to engaging in a competitive match.) Similarly, the End-Sharing Account gets **Competitive Tennis** wrong but gets **Surprise Party** *right.* (Your sister-in-law may not be able to consent to your lying to her about the party, but she can share your end.)

Here is how Kerstein combines the two accounts:

> **Hybrid Account:** S uses P merely as a means if it is reasonable for S to believe both that (a) P is unable to consent to S using him, and (b) P cannot share the proximate ends that S is pursuing in using P.

According to the **Hybrid Account** if we wish to avoid using another as a mere means we must either act in ways that we reasonably believe they can consent to or pursue ends that we reasonably believe they can share. If we fail at both of these, then we are using them as a mere means.

Does the Hybrid Account capture the heart of what we care about when we care about not treating others as mere means? Earlier we suggested that what is attractive about the formula of humanity is that it captures the intuitive idea that we shouldn't treat persons as mere objects or instruments for our use. Insofar as we are inclined to think that respect for persons is the heart or foundation of our moral principles, what this idea presumably amounts to is respect for that aspect of us that makes us the kinds of beings that we are: the ability to set our own ends and guide our conduct in accordance with reasons. Respect for persons, so understood, is a respect for this aspect of who we are. For anyone who buys into this conception of morality, an adequate account of the formula of humanity ought to reflect this conception.

Does the Hybrid Account do this? I think the answer is yes. There are two aspects to the Hybrid Account: consent and end-sharing. Consider the second of these. According to the Hybrid Account, if in interacting with someone we have reason to believe that they can share our ends, then we are not treating them as mere means. The emphasis in recognizing the ends of the other person is important. By recognizing that the other person has ends of their own, we are recognizing their ability to set their own ends and respecting this capacity. The end-sharing side of the Hybrid Account captures the insight that a person's end-setting capacity is central to their status as being deserving of respect.

Consider now the consent side of the Hybrid Account. We do not treat others as mere means if it is reasonable for us to believe that we are treating them in ways that they can consent to. This side of the Hybrid Account doesn't focus on our capacity as end-setters, but rather our capacity to guide our behavior by the light of reason. In order to treat someone in ways that they can consent to, we have to offer them the opportunity to dissent. In offering them this opportunity we typically have to describe to them our end, and the reasons that support the pursuit of this end. In so doing we are providing them with the justification for our action as we understand it, and leaving it up to them whether to accept the course of action or not. In doing this, we are respecting the person's ability to guide their behavior by the light of reason. To do so is to treat someone as an agent rather than a mere object.

I think these considerations suggest strongly that Hybrid Account captures the core insight of the formula of humanity. Whether Kerstein's Hybrid Account is the *best*

way to unpack the formula of humanity, I do not at this time know. A full investigation is outside the scope of the current project, but I urge interested readers to read Kerstein's book on the topic.

5.4 The Hybrid Account and the Formula of Humanity

The thesis of this chapter is that the moral law, understood as the formula of humanity, is a constitutive principle of shared agency. If we wish to develop this thesis – or assess the plausibility of it – we need to unpack what it means. My aim in the current section has been to do just that. I have suggested that we unpack the formula of humanity in the way that Samuel Kerstein does in his book *How to Treat Persons*, that is, by way of the Hybrid Account:

> **Hybrid Account**: S uses P merely as a means if it is reasonable for S to believe both that (a) P is unable to consent to S using him, and (b) P cannot share the proximate ends that S is pursuing in using P.

The formula of humanity understood in terms of the hybrid account, can be understood as follows:

> **Formula of Humanity**: Never treat others in such a way that it is reasonable to believe both that (a) they are unable to consent to such treatment, and (b) they cannot share the proximate ends that such treatment involves.

The *Shared Agency Hypothesis* holds that the formula so understood is a constitutive principle of shared agency. The rest of this chapter will develop and defend this idea.

6. Gilbert's Plural Subject Theory

6.1 The Details

Up to this point I have been considering the Shared Agency Hypothesis in isolation of theories of shared agency. The rest of this chapter looks to theories of shared agency to support this hypothesis. I will focus my attentions on two such views: Margaret Gilbert's "Plural Subject Theory" and Michael Bratman's "Planning Theory". What we will find in this investigation is that, while these theories come very close to explicitly rejecting the Shared Agency Hypothesis, their views will in fact supply us with the resources to defend it. In this section I focus on Gilbert's view, in the next I concern myself with Bratman's.

Much of the theory of human action is focused on analyzing the actions of individuals (and the components thereof), or what I will call *singular actions.* But much of human agency is exercised in conjunction with others, whether it's playing a game of basketball, writing new tax policy, performing a song live, building a shed, walking together, or dancing the tango. Each of these collective activities involve more than just individual agents converging accidentally on a common goal, as when you and I, unbeknownst to the other, pick up trash on the beach. When this happens, we are each performing singular actions in a common space, with a common goal, but we don't perform an action *together* – our agency is not *shared*. However, upon meeting each other while cleaning up the beach that day, we may decide that next weekend we shall clean the beach *together*. When we do so will be performing a *collective* or *shared* action. The central task of theories of collective action is to explain what the difference is between cleaning up the beach in a common space and cleaning up the beach *together*.

How do we get from agents acting individually to agents acting collectively? How do I get from dancing, individually, to dancing with a partner, collectively? How do I get from playing my instrument while others play theirs, to us playing our instruments together? How do I get from cleaning up the beach while you do as well, to us cleaning up the beach *together*?

Theories of collective action can be distinguished according to how they answer a few initial ontological questions. Is a collective action something over and above the individual actions that make it up? Theorists who provide a positive answer to this question think that group agency or group intention is an irreducible phenomenon that does not reside wholly in the head of any one agent. Gilbert (1989, 2000) calls these views *holistic accounts*. Theorists who reject holism – sometimes called individualists – hold that the phenomenon of collective action can be fully explained without appealing to *sui generis* "units" or "plural subjects", but rather by appealing only to the "mental economy" inside the minds of individual agents.

Our second axis by which theories of collective action can be compared concerns whether the explanans of shared agency are reducible to the explanans of individual agency. (For example, it is common to explain singular agency by reference to the agent's intentions. Can we explain shared agency in terms of the intentions of individual agents?) Since holists hold that collective agency is something over and above the agency of the individuals that comprise it, holism is by definition a non-reductive theory. However, this does not entail that all individualist views are reductive. Some views, such as John Searle's (1990), holds that collective actions are explained by attitudes existing in

the mind of each party to the collective act (individualism), but these attitudes are irreducibly collective ("we"-intentions, he calls them). Most individualist views, though, are reductivist, since the philosophers who posit such views are often driven by concerns of parsimony. The challenge they tend to task themselves with is to account for shared agency without positing anything beyond that required for singular agency.

The third axis against which we can compare theories of collective action concerns how robust the accounts are: does the theory require only a minimal analysis appealing to only the familiar mental states of singular action, or does it require a complicated analysis and the positing of sui generis mental states or plural subjects? I will call theories of the former kind minimalist theories and theories of the later kind robust theories. Those theories that fall into the middle I will call moderate theories.

The first view I will sketch – "Gilbert's Plural Subject View" – is a holistic, non-reductive, robust view. As we will see, she argues that in order to account for shared agency we need to posit *plural subjects* to whom collective actions are attributed. You and I clean the beach *together* when we form a plural subject to do so. These plural subjects are holistic – they stand over and above the attitudes in your and my head – and since they are holistic, plural subjects are also irreducible. The robustness of Gilbert's view is not just due to its holism, but also to the cluster of rights and obligations that she places at the center of shared agency.

The second view I will sketch – Bratman's "Planning Theory" – is individualist, reductive, and moderate. On Bratman's view, two or more people act together when their intentions mesh in the right way. You and I clean the beach together when we have

interlocking plans to do so. Bratman's Planning Theory is individualist because it grounds shared agency fully within the "heads" of individual agents, it is reductive because the attitudes that give rise to shared agency are the same kinds of attitudes that comprise singular agency, and it is moderate rather than robust because it does not posit *sui generis* mental states or plural entities. The Planning Theory is moderate rather than minimalist because it relies on the attitudes and norms of "planning agency", which some action theorists think is a robust way to account for singular agency. In addition, it involves a complicated account of when the intentions of two or more agents "mesh", a crucial step in his account for creating a shared intention.

Now that we have the conceptual space of theories of shared agency mapped out, let's turn to Gilbert's Plural Subject Theory. According to this view, two or more agents act together by forming a plural subject. And two or more agents form a plural subject, when and only when, they form a joint commitment. To understand her view we need to understand two things. First, what is a plural subject? And second, what is a joint commitment?

In answering these questions let's start where Gilbert often starts: the core data points that give rise to "three central criteria of adequacy for an account of shared intention" (2009, 171). We can come to recognize the phenomena by reflecting on common, recognizable cases of shared agency.

Our first phenomenon will correspond to what Gilbert calls the Disjunction Criterion. Consider the following case, quoted in full from Gilbert (2009):

> **Olive's Report**: Our plan was to hike to the top of the hill. We arrived at the hill and started up. As he told me later, Ned realized early on that it would be too

> much for him to go all the way to the top and decided that he would only go half way. Though he no longer had any intention of hiking to the top of the hill, he had as yet said nothing about this to me, thinking it best to wait until we were at least half way up before doing so. Before then we encountered Pam who asked me how far we intended to go. I said that our intention was to hike to the top of the hill, as indeed it was. (p. 171-172)

In **Olive's Report** Ned and Olive agree to go hiking together up a steep hill. As they set off, it seems accurate to describe them as *hiking up the mountain together* and having the intention to do so. However, early on in the hike Ned decides that he won't be able to make it up to the top. When Olive, after this, tells Pam that they intend to go to the top, is her assertion truthful? Gilbert thinks it is. "[P]eople may share an intention though at least one of them lacks a personal contributory intention" (ibid 172). A personal intention is an intention ascribable to an individual which we would express in the form of "I intend to ϕ". Gilbert distinguishes personal intentions from those intentions we ascribe to groups or collectives, such as the intention that we might ascribe to the duo, Olive and Ned. Gilbert calls these kinds of intentions *shared intentions*. A personal *contributory* intention is a personal intention to *contribute* one's part toward the goal of a shared intention. One might think that a collective action or group intention requires that all participants have a personal contributory intention concerning doing one's part to satisfy the shared intention, but **Olive's Report** suggests this is not the case. In fact, Gilbert seems to think that it's possible for *no party* to a shared intention to have a personal contributory intention:

> I would not take Olive not to involve herself in an inconsistency if she went on: "As it happens, when we met Pam, I was in the same position as Ned: I'd decided that I would not go all the way to the top of the hill, though I hadn't yet broached the subject with him." (ibid)

Gilbert thinks that the lesson can be broadened to include, not just personal contributory intentions, but also what she calls *correlative personal intentions*, which is "any personal intention geared specifically to the satisfaction of a given shared intention" (ibid). Whereas a personal contributory intention is an intention to contribute one's part in a shared act, a correlative personal intention is *any* intention which concerns the collective act one is a part of. So, for instance, I might intend *that we climb a mountain*. This isn't a contributory intention *per se* (though this intention might in turn *commit* me to certain contributory intentions) but rather a *correlative* personal intention.[25]

The lesson that Gilbert thinks we learn from cases such as **Olive's Report** is the following:

> **Disjunction Criterion:** an adequate account of shared intention is such that it is not necessarily the case that for every shared intention, on that account, there be correlative personal intentions of the individual parties. (ibid)

Gilbert notes that, as far as the Disjunction Criteria is concerned, it is possible that correlative personal intentions can be *sufficient* for shared intention. We will shortly see, however, that Gilbert thinks that this is not the case

The second data point concerns when a shared intention can be changed or rescinded. The case that will illustrate this concerns two people, Rom and Queenie, who are walking together to go shopping:

> **Slow Walker** Queenie's pace begins to slow. In a tone of mild rebuke Rom says "Can you hurry up a bit? We won't be able to get any shopping done at this rate! Queenie says "Sorry!" and starts to move more quickly. Later she stops and for some reason announces: "That's it! I'm not going any further!" (ibid, p. 173)

[25] As we will see shortly, Bratman holds that correlative personal intentions of this form are the building blocks of shared agency.

In **Slow Walker** Queenie and Rom intend to walk to the shopping mall *together*. How is Rom likely to react to Queenie when she suddenly decides to stop walking? Gilbert thinks that "Rom is likely to be taken aback" (ibid). Further, she thinks that Rom would be justified in rebuking Queenie for doing something which Queenie is not entitled to do, namely unilaterally deciding to bow out of the shared intention. From this, Gilbert draws the following:

> **Concurrence Criterion**: absent special background understandings, the concurrence of all parties is required in order that a given shared intention be changed or rescinded, or that a given party be released from participating in it. (ibid)

The **Concurrence Criterion** and the **Disjunction Criterion** are related to an interesting feature of shared intentions: an individual cannot unilaterally rescind or change a shared intention. In the Concurrence Criterion this feature results in the idea that all parties to a shared intention must concur if an intention is to be changed or rescinded. In the **Disjunction Criterion** this feature results in the fact that a shared intention can persist absent personal correlative intentions.

Our final criterion is closely related to the previous two. Recall that Gilbert thinks that Rom is justified in rebuking Queenie for unilaterally trying to rescind the shared intention. What gives Rom standing to do this? Gilbert thinks it comes from the nature of a shared intention:

> **Obligation Criterion**: each party to a shared intention is obligated to each to act as appropriate to the shared intention in conjunction with the rest. (ibid, p. 175)

Note here that the obligations are *to* the other parties of the shared intention. These obligations are the corollary of what Abraham Roth calls *contralateral commitments*, the

personal commitments that each party to a shared intention has to the other parties to do their part in achieving the goal contained in the intention.[26] The idea behind the **Obligation Criterion** is that when you and I form a shared intention to do something, in forming that intention we have obligated ourselves to each other. This obligation is what explains the standing that you would have to rebuke me were I to decide, unilaterally, to not do my part in our collective act.

Gilbert argues that her Plural Subject Theory can explain these three criteria. Our next stop on the road to understanding the view is understanding why Gilbert thinks that a theory which relies only on personal intentions cannot account for the Obligation, Disjunction, and Concurrence Criteria. According to Gilbert, personal intentions create what she calls "personal commitments" (2009). A personal commitment, on her view, is a combination of what I call volitional and consistency commitments (see chapter 2). Consider the following discussion from her (2003):

> Suppose… that Janice decides to have lunch at Café Earth today. I take it that she is now in some sense committed to having lunch at Café Earth today. She can, of course, change her mind. But as long as she does not do so, she is committed.
>
> A personal decision creates what I call a personal commitment. By this I mean a commitment of a person A, such that A is in a position unilaterally to make and unilaterally to unmake or rescind it. …
>
> A commitment has what philosophers refer to as *normative force*. If one violates a commitment to which one is subject, one has done what in some sense one was not supposed to do. One has to some extent and in some sense done something wrong – something open to criticism. (p. 47)

In this passage, Gilbert identifies two aspects of a personal commitment, a volitional aspect and a normative aspect. The volitional aspect of the commitment is identical to

[26] Contralateral commitments will be discussed in further detail later in the section.

what I have been calling a volitional commitment. When you make this kind of a commitment, your agency becomes structured toward taking the means to achieve the end. So, if Janice is committed to having lunch at Café Earth, she will at some point begin to take the means toward having lunch at Café Earth. What Gilbert is calling the "normative" aspect of a commitment is roughly what I have been calling a consistency commitment. If Janice never leaves her desk, she has gone wrong vis-à-vis her commitment. Elsewhere, Gilbert identifies the normative aspect of a personal commitment as "a matter of what might be referred to as a *rational requirement*" (2009 p. 179). Janice, in being means-end incoherent, acts irrationally.

For Gilbert, an important aspect of a personal commitment is that they can be created and rescinded unilaterally. To create a commitment, all I need to do is form an intention. And to rescind an intention, all I need to do is decide to no longer go through with my prior intention.[27]

It is precisely this aspect of personal commitments which Gilbert thinks make them unable to serve as the foundation of shared agency. A commitment which can be unilaterally created and rescinded will not be able to explain the Concurrence Criterion, the Disjunction Criterion, or the Obligation Criterion. If a shared intention was the

[27] One might disagree that we really have the same unilateral power to rescind commitments that we do to create them. Michael Bratman (1987, 2007) famously argued that one of the functional features of intentions is their stability. If we can rescind them willy nilly, then they will lack this important feature which itself contributes to the normative role they play in our practical reasoning. I take it that Gilbert's point is not that commitments lack this feature, but rather that I *can* rescind my personal commitments without appealing to anyone else – I, and only I, have the authority to rescind my personal commitments. This doesn't mean that there are *no* normative pressures pushing against this authority we possess. In fact, we might think that, all else being equal, I (normatively speaking) can only rescind a personal commitment if I have sufficient reason to do so.

concatenation of personal intentions, then contra the Disjunction Criterion, one could dissolve the shared intention by rescinding one's personal intention. According to the Concurrence Criterion rescinding a shared intention requires concurrence from all parties, but an individual does not need concurrence from a third party to rescind a personal intention. Finally, personal intentions, in themselves, do not create obligations toward others. That is, as far as the norms associated with our personal intentions are concerned, we would have no standing to rebuke each other if one of us were to fail to follow through on our intentions.

Gilbert thinks that where personal commitments failed, *joint commitments* can succeed in accounting for the normative features of shared agency. In brief, "a joint commitment is [a] commitment of two or more people" (2003, p. 49). Joint commitments are *holistic* in the sense that they are irreducible – they cannot be reduced to personal commitments. Gilbert calls a joint commitment's holism its "core feature" (ibid). The following are the other central features of a joint commitment:

1. Each party in a joint commitment is answerable to the other parties

2. Joint commitments are created by participation of *all* parties; "a joint commitment cannot be created by a single party acting unilaterally" (2014, p. 40).

3. One cannot unilaterally rescind a joint commitment; rather, joint commitments can be rescinded "only by the parties together" (2003, p. 50)

4. Joint commitments give rise to dependent individual commitments. These are commitments of the parties to do their part in the joint intention. (These are not personal commitments because the individuals do not have the unilateral authority to rescind them.) These commitments come into existence when a joint commitment is formed and can only be rescinded if the joint commitment is rescinded.

5. The content of a joint commitment is for the collective body to act in a certain way. If you and I have a joint commitment to dance the tango, we have made a commitment to dance together in a certain way.

Joint Commitments, so understood, satisfy our three criteria. Because joint commitments are holistic, their persistence is not dependent on the persistence of individual commitments, and so rescinding the latter does not result in rescinding the former (Disjunction Criterion). Because joint commitments persist even when personal corelative commitments are rescinded, the only way a joint commitment can be rescinded is if all the parties agree to do so (Concurrence Criterion). And finally, because joint commitments can be rescinded only if all the parties agree to do so, joint commitments obligate the parties to each other (Obligation Criterion).

So we know what a joint commitment is, but how do they get formed? According to Gilbert, parties must openly express "a readiness to be jointly committed" in conditions of common knowledge (2003, 53). Perhaps surprisingly, Gilbert think that "this is pretty much the whole story regarding the creation of … joint commitments" (2014, p. 48). So these necessary conditions are also jointly sufficient for forming a joint commitment. All that the potential parties to a joint commitment need to do in order to form a joint commitment, then, is express their respective readiness to do so.

Perhaps an example will help. José wants to go see the new Marvel movie with Jane. He sends a text message to her:

José: Want to go see *Black Panther*?

Jane: Yes! What time?

José: How about the 5:20 showing?

Jane: I can do that!

José: I'll pick you up at 4:45

Jane: Sounds good. See you then.

In this exchange of text messages, José and Jane both express a readiness to be jointly committed. José does so in his initial text message, Jane conditionally so in her response (i.e., she expresses a readiness conditional on the movie showtime being an appropriate one) and categorically so in her follow up ("I can do that!"). At this time José and Jane are jointly committed to seeing *Black Panther*, that is, they share an intention to do as such. If José does not show up to pick up Jane, she would be warranted in being upset, and if Jane is not ready for Jose, he would be entitled to be upset (obligation criterion). Finally, neither Jane nor José can unilaterally decide to see another movie instead of *Black Panther* (concurrence criterion).

The expression of readiness that one offers when one enters into a joint commitment can be made verbally, via messages (as in our example), or even with a nod or wink.

6.2 Plural Subject Theory and the Shared Agency Hypothesis

I will now turn to my defense of the Shared Agency Hypothesis. I will argue that in order for parties to enter into a joint commitment with each other, they must treat each other as ends in themselves. I will try to show that the resources for defending the Shared Agency Hypothesis can be found within Gilbert's Plural Subject View.

My argument begins by inquiring into what is involved in an expression of readiness. In particular, *what is it to express readiness*? Unfortunately, I haven't come across any text where Glibert considers this question. In fact, she states that "[i]t is not clear that there is any very helpful way of breaking down the notion of expressing one's readiness to be jointly committed" (2014, p. 48). But I believe that we can break this notion down, or at least give some content to it, if not fully analyze it.

I take it that when Gilbert says that expressions of joint readiness by all the parties to a joint action is sufficient she means not just the *expression*, but also the actual *readiness*. That is, all parties must be willing, and they must express to each other that this is so. Expressing readiness is like expressing gratitude. One can't express gratitude unless one really has it. Of course, one can make an expression *as if* they were grateful – I can say thank you, but not really mean it – but when one is doing this, one is not really *expressing gratitude.* The same goes for expressing one's readiness. One can say, "Let's go to a movie" but if one doesn't mean it – if one isn't genuinely willing – than this utterance is not an expression of *readiness*, but just a misleading expression, and in making such an expression, one hasn't entered into a joint commitment.

This underlying idea is important. In order to enter into a joint commitment, one's expression of readiness must *express one's readiness*. In other words, the expression must reflect an underlying willingness on the part of the agent to actually enter into a joint commitment.[28] An insincere or inauthentic expression will not suffice.

[28] Gilbert says that she "prefer[s] the term 'readiness' [because it is] less suggestive of a strong form of voluntariness" than the term "willingness" is (2014, p. 47). I use 'willingness' here only for stylistic reasons. But this concern of Gilbert's will be interrogated shortly.

One might be inclined to think that what I have just said is false. Imagine that in our previous example José was being insincere with Jane. When he asked Jane to go to the movie, he wasn't really sincere – perhaps he was playing a prank on her. That is, when José inevitably doesn't show up, in such a case we would tend to judge that Jane is entitled to rebuke José. But if this is the case, we might think that José has indeed formed a joint commitment with Jane, for her entitlement to rebuke José might seem to be the same phenomena that the obligation criterion tracks, and so a joint commitment must be appealed to in order to explain it.

I think this line of thought is mistaken. It is important to keep in mind that the obligations and entitlements that Gilbert identifies as constitutive of joint commitments are not *moral* obligations or entitlements. Rather, these obligations and entitlements are closer to what we would call rational requirements than those of the moral variety. But I would suggest that the intuition we have about Jane's entitlement to rebuke José when he doesn't come pick her up is not an entitlement associated with a joint commitment, but rather a *moral entitlement*. José has used Jane as a mere means, an object for his amusement. She cannot consent to this treatment (for José is engaged in a form of deception), and she cannot share his ends (since his end – to prank Jane – relies on her not sharing his end).[29] The entitlement that Jane has to rebuke José is the entitlement associated with the fact that José has used her as a mere means, not the entitlement associated with a joint commitment.

[29] We should not lose sight of the fact that the thesis of this chapter is that the moral law is a constitutive principle of shared agency, and that we have just arrived at yet another case where treating someone as a mere means prevents the formation of shared agency.

Now, what is involved with this readiness that Gilbert says is at the heart of a joint commitment? The best way to answer this question is to turn straightaway to an objection to the Shared Agency Hypothesis found in Gilbert's oeuvre, namely that coercion is not an impediment into entering joint commitments. It is in her articulation of this claim that we find her best articulation of what might be at the bottom of an agent's "readiness" to forming a joint commitment. My subsequent response to this argument – which I will be calling *Gilbert's Coercion Argument* – will constitute the final step in my argument that the moral law is a constitutive principle of shared agency, where shared agency is understood in terms of the *Plural Subject Account*.

6.3 Gilbert's Coercion Argument

Properly speaking, Gilbert's Coercion Argument concerns the relationship between coercion and agreements. Gilbert wishes to argue against the commonly received view according to which a coerced agreement is not an agreement. On Gilbert's view an agreement, like a shared intention, is a kind of *joint commitment*. As we will see, Gilbert's Coercion Argument is really an argument that an individual A can coerce an individual B into entering into a joint commitment with her. If this argument is successful, it will quite obviously undermine the Shared Agency Hypothesis.

Gilbert identifies two distinct kinds of arguments that have been offered for the conclusion that coerced agreements are not agreements of any kind. Gilbert calls these "the obligation argument" and the "voluntariness argument." Let's consider them both.

The Obligation Argument

1. Any genuine agreement generates a moral obligation to abide by it: this is a conceptual matter.
2. If any apparent agreement is made in the face of coercion, there is then no moral obligation to abide by it.
3. Therefore, a genuine agreement cannot be made in the face of coercion. (1996b, p. 282)

Gilbert argues that the **Obligation Argument** is inconsistent with what she calls "the knowledge of agreements assumption": "if someone enters into an agreement, then she knows that she does" (ibid, 284).[30] If both the Obligation Argument and the knowledge of agreements assumption where true, then in order to make an agreement one would have to make judgments about the presence or absence of coercion, which can be a tricky thing to do in some circumstances. But, Gilbert reasons, determining whether one has made an agreement is not tricky, so agreements must not generate moral obligations. And if agreements don't generate moral obligations, then the conclusion of the Obligation Argument does not follow.

While not much hangs on this for my view, I find the knowledge of agreements assumption to be implausible. It is not uncommon to be unsure of whether one has made an agreement when one in fact has, or even to believe that one has not when one in fact *has*. Consider the following case:

> **Spacey Steve:** Steve had a tough day at work. When he comes home his daughter Betty asks if Steve can take her to baseball practice later that evening. Steve says yes, but while he is talking with her he is also thinking of his day at work. He walks into the kitchen to grab a cold beer to help him relax, not realizing that he has just agreed to take Betty to work. Steve has made an agreement, but does not know that he does.

[30] Gilbert stresses that "this assumption concerns informal agreements as these are conceived of in everyday life" (ibid).

Is the story of Spacey Steve incoherent? If the knowledge of agreements assumption is correct, it must be. But it seems perfectly coherent, in fact it likely rings true to many of our experiences with our parents. What is likely going in cases like Spacey Steve is that an individual's second order awareness is not tracking their first-order awareness. The same thing happens when we engage in "blind driving" such as when we hop on the freeway to get to work when we really intend to go to the store to pick up some milk. We can be driving to work completely unaware that we are doing so often because our second-order awareness is preoccupied. The problem with Gilbert's knowledge of agreement's assumption is that it requires second-order awareness of an agreement in order for an agreement to be made. But it's unclear why this should be.

I now turn to the second argument Gilbert considers in favor of the conclusion that there cannot be coerced agreements. This argument is the more relevant of the two for my purposes.

The Voluntariness Argument

1. Agreements are by definition voluntary: they cannot come about against the will of either party.
2. Coercion does not allow an agreement to be voluntary.
3. Therefore, coercion rules out the very possibility of agreement. (1996b, p. 286)

This argument, it should be noted, looks quite similar to the argument I advance in favor of the Shared Agency Hypothesis. Gilbert thinks that this argument equivocates on 'voluntariness' in premises 1 and 2. Gilbert asks us to first reflect on the following case:

> **Crafty Trickster** Betty has decided not to buy a certain house. However, a crafty trickster gets her to sign her name at the bottom of an agreement to purchase it

> without realizing what she is doing. Perhaps he covers the agreement in such a way that she thinks she is simply giving him her autograph. (ibid, pp. 286-287)

In **Crafty Trickster**, Betty literally signs a contract, but it seems wrong to say that she has genuinely made an *agreement* to buy the house, as she thought that she was giving an autograph, not signing a contract.[31] According to Gilbert, the sense in which this was an agreement against Betty's will is in what she calls the 'decision-for' sense: "she never made any decision in favor of signing the agreement in question" (p. 287).

Gilbert asks us to compare **Crafty Trickster** with a case like this:

> **Forced Signing** Betty has decided not to buy a certain house. However, a mobster gets her to sign her name at the bottom of an agreement by pointing a gun to her head and threatening to shoot if she doesn't sign. Because of his threat, she signs the agreement.

In **Forced Signing**, the sense in which Betty's signing is against her will is not in the decision-for sense. Gilbert notes that Betty in this case was presented with a choice: sign the contract or get shot. She chose to sign the contract. So unlike in **Crafty Trickster**, there is a sense in which Betty has done something voluntarily. While her decision may have been coerced, it was still a decision and to that extent voluntary.[32]

Using this distinction between acting against one's will in the decision-for sense and acting against one's will when coerced to do so, Gilbert argues that the Voluntariness argument equivocates on two different senses of 'voluntary': premise 1 makes use of the decision-for sense of 'voluntary', while premise 2 uses 'voluntary' in the coercion-free

---

[31] We should note, yet again, that, while this is not a case of potential shared agency, we have a case where using someone as a mere means has interfered with the formation of a joint-commitment. This, of course, is case of deception.

[32] The point that Gilbert is making here is similar to the point that Ayer makes in his famous essay "Freedom and Necessity" regarding the fact that the presence of coercion does not exclude the presence of an ability to do otherwise.

sense. If 'voluntary' is used in these two different ways, then obviously the argument's conclusion does not follow.

Is Gilbert right that agreements – and by extension joint commitments – do not need to be free of coercion in order for them to be genuine agreements? Did Betty genuinely *agree* to buy the house when she was held at gunpoint? I think Gilbert's analysis of this case is wrong.

Gilbert's line of argument is not uncommon in the history of philosophy. As Anderson (2017) points out, Aquinas and Hobbes, like Gilbert, also drew a distinction between coercion and threats. For both of these thinkers, coercion involves forcibly moving another's body in such a way that you remove their ability to do otherwise. When this happens, the person's will plays no part in the action, and their movements are involuntarily. A threat of violence, however, does no such thing. Providing a threat of violence presents one with a different set of reasons for action, and so changes the rational calculus concerning what should be done.

I think this distinction perverts practical reasoning. Threatening someone with violence forces them into a normative landscape with only one real option. In most circumstances, avoiding violence is the choice that any person would take, and this is precisely why the coercer opts to threaten violence. While it is true that there is a sense in which one exercises one's agency when one is complying with the orders of a coercer, that there was a choice in the matter is nothing more than a mere metaphysical possibility. Focusing on the fact that non-compliance is metaphysically possible discounts the sense in which non-compliance in most such situations is a normative

impossibility for many people.[33] In **Forced Signing**, does Betty really experience non-compliance as a normative possibility? If Betty is like me or you, the answer to this question is surely no. If her only other option is to be shot, she will see only one path she can take, and she will take this path. To say that no one has forced her to take this path, that she is voluntarily doing so, is absurd. By threatening Betty with violence, the gunman has forced her into a normative context with only one exit: signing the document.

Gilbert might insist that she agrees to everything I have said, and that it is precisely for this reason that Betty would not be *morally responsible* for signing the document and probably not morally bound to it. But the question of moral responsibility and moral obligation misses the point. An agreement between two parties, if it is to *be* an agreement, must reflect the practical reasoning of each party to the agreement – the terms of the agreement must be *endorsed* by each party. But being coerced to accept an agreement by threat of violence binds me to the terms of the agreement only by continued fear of violence. Gilbert disagrees. She thinks that we have *pro tanto* obligations to keep our agreements, even when those agreements are reached via coercion. "The obligations of agreement stand as long as the corresponding joint commitment does. Whether or not the commitments stands is a matter for the participants to decide" (p. 299).

Here I think Gilbert gets into trouble. Recall how she holds, plausibly in my view, that joint commitments cannot be created unilaterally – all participants to a joint

[33] I say *most* circumstances because sometimes the violence one would have to endure for not complying is less worse than what compliance requires. Sometimes death is the easier, softer way.

commitment must participate in its creation. The problem is that if Gilbert is right about how to interpret cases like **Forced Signing**, then individuals *can* unilaterally create joint commitments, namely by threatening violence for non-compliance on those who they wish to form such commitments with. So Gilbert either must give up how she interprets **Forced Signing** or she must give up the claim that joint commitments cannot be created unilaterally. Gilbert should give up the former.

This is an important point, and worth dwelling on. Why think that joint commitments cannot be created unilaterally.[34] There are, I think, two reasons for accepting this claim. First, joint commitments are *commitments*. While they may not be reducible to individual personal commitments, they do reflect the volitional structure of the parties to the commitment, and in reflecting the volitional structure of the parties, they must reflect where the individuals stands vis-à-vis the content of the commitment. You cannot unilaterally create a joint commitment that binds me because you cannot unilaterally exercise *my* agency. The second reason we should accept the claim that joint commitments cannot be created unilaterally is a related point that concerns the normative structure of joint commitments. As we've seen, joint commitments give rise to obligations. These obligations, I suggested, are to shared agency, what the instrumental principle is to singular agency. In chapter 2 I argued that the normativity of any given instance of the instrumental principle is generated by one's commitment to engage in the activity the principle is constitutive of. The same can be said for the normativity of the

[34] Gilbert doesn't offer much of a rational here, so what I am about to offer in defense of the claim goes beyond her view and makes use of my Commitment View (see chapter 2).

obligations constitutive of shared agency. The source of this normativity comes from the joint commitment, and it binds to one's will in virtue of the joint commitment being an expression of a person's will (much as a singular commitment is an expression of a person's will). If you could force me into a joint commitment, you could force these obligations on me. But you can't force these obligations on me, for you can't reach into my will and commit me to the activity. Only I can do that. But this, I have been arguing, is *exactly* what a case of coercion is. It is one person trying to reach into the will of another.

Now, Gilbert might remind us at this moment that the obligations created by a joint obligation are just *pro-tanto* obligations. In Betty's case, they will likely be outweighed when they conflict with other obligations. But it's not clear to me why we should think they are present *at all*, regardless of how much they weigh. Is there *any* reason to think that Betty's coercer would be entitled to (non-morally) rebuke Betty if she failed to comply with terms of the "agreement"? Is there any reason to think that Betty should (non-morally) rebuke *herself* if she failed to comply? Insofar as coercion is practically necessary[35] to get one to perform a certain action, then the only normative requirement at play in this case concerns the prudential reason associated with avoiding violence done against one. But of course, this reason has nothing to do with the creation of a joint-commitment.

[35] By "practically necessary" I mean that being coerced was the primary reason that one behaved in a certain way.

I have been arguing that an individual A cannot coerce an individual B into forming a joint commitment. This can't happen because it is inconsistent with the very nature of a joint commitment. It would involve the unilateral creation of a joint commitment on the part of A, as well as A imposing normative requirements on B through the use of force. We have also seen, as in the case of **Crafty Trickster** that deception is inconsistent with the formation of a joint commitment. If I am unaware of the ends to which my agency is being put to use, then I do not have a commitment to bring about this end, and so am not part of a joint commitment to seeing to it that this end be achieved.

Both of these cases involve an individual being used merely as a means, and we have found that the individual so being used is not a party to a joint commitment as a result. Can we make a general argument for the Shared Agency Hypothesis? I think we can. Let me first provide you with a rational, and then with a more formal argument.

In "Agreements, Coercion, and Obligation", Gilbert provides a bit more of a gloss on the idea of "readiness" than she does elsewhere in her oeuvre: "I use this term to indicate only that a state of the agent's will is at issue or, more specifically, a positive inclination of the agent's will" (1996b, p. 285).[36] Gilbert clearly thinks that the formation of a joint commitment involves an exercise of agency, and that one is bound by a joint commitment only if one's agency is implicated. The reason the moral law is a constitutive principle of shared agency is that when A uses B as a mere means in the

[36] In this paper, Gilbert uses the term "willingness" rather than readiness, but she notes, in the very next sentence, that "The terms 'readiness' or 'preparedness' might have been used instead" (ibid).

pursuit of end E, B's agency is not implicated in the appropriate way. B is not obligated to A to perform the means required for achieving E, so B does not form a joint commitment with A.

1. If A and B form a joint commitment to pursue E, then A is obligated to B to take means M_A and B is obligated to A to take means M_B. [Obligation Criterion]

2. If A uses B as a mere means in the pursuit of E, then B is not obligated to A to take means M_B.

3. So, if A uses B as a mere means in the pursuit of E, then A and B do not form a joint commitment.

Allow me to provide a justification for premise 2. Remember that the kind of obligation in question is not a moral obligation, but rather an obligation similar to the rational requirements of instrumental rationality. The source of such requirements lie in the will of the individual who falls within the scope of the requirement. I obligate myself to take the means because I commit to taking the end.[37] That my will is implicated is essential. When A treats me as a mere means in the pursuit of an end, my volition is not implicated in the right way because I am being treated in a way that I can't consent to, and because I cannot share the end that A is pursuing.[38] As a result I am not committing to the end, and so I do not become obligated to take the means. And if I do not become obligated to take the means, then I cannot be a party to the joint commitment. This is why the moral law is constitutive of shared agency, at least if we accept the Plural Subject View.

[37] See chapter 2.

[38] Here I am invoking the Hybrid view

6.4 Interlude: Roth's Contralateral Commitments

We've seen that Margaret Gilbert identifies a normative structure that is constitutive of shared agency. The structure is important to my defense of the Shared Agency Hypothesis, for the lynchpin of that argument is the claim that one cannot unilaterally impose the obligations constitutive of shared agency onto another, for the source of their normativity must ultimately rest in the commitments of those who are bound by them. If it turns out that such obligations are not constitutive of shared agency, there will be a giant lacuna in my argument. As we will see, Michael Bratman is a theorist who claims just this. So before turning to Bratman's view, I'd like to briefly discuss another theorist who identifies a similar normative structure to shared agency. Ultimately, I will argue that Bratman should accept the normative structure Gilbert recognizes – that it is in fact consistent with other elements of his so-called "Planning Theory" of agency.

The primary thesis that Roth defends across several papers (2003, 2004, 2014) is what he sometimes refers to as "**practical intimacy**", which is the idea that one can act directly on the intentions issued by another. According to Roth, shared agency is constructed out of intentions so shared. I won't be discussing the idea of acting on another's intentions – this will take me too far afield – but we will be looking into his motivation for adopting this thesis, as it bears on our more immediate concern: the normative dimensions of shared agency that Gilbert calls the "obligations" and "entitlements" of shared agency.[39]

[39] In what follows I will largely by relying on Roth (2004).

Roth argues that every participant to a shared action has three commitments: a participatory commitment to achieving the goal of the shared action, an ipsilateral commitment to themselves (to adequately discharge their intention by taking the required means to achieving the goal), and a *contralateral* commitment to each participant in the shared action to do their share.

Now, participatory commitments and ipsilateral commitments come with any kind of intention. These correspond roughly to what I have been calling *volitional commitments* and *consistency commitments*, respectively. A participatory commitment is the kind of commitment one makes toward achieving the end contained in an intention. An ipsilateral commitment is a commitment one makes to oneself to what is required to fulfill one's intention. But why do we need contralateral commitments? An example will help answer the question. Imagine that Lisa and Zac are on a walk together. Were you to ask them what they are doing, they would both say, "we're walking together." This is a paradigmatic shared action. Now imagine that Zac begins to walk very fast. If this were to result, Lisa would be entitled to shout, "slow down" or "hey! I thought we were walking together!"

This example of two people walking together is of course a famous example that Gilbert uses.[40] Roth suggests that we should understand the normative framework here as a web of commitments. Zac has a commitment *to Lisa* to walk at a measured pace, and Lisa has a commitment *to Zac* not to walk too slowly. Presumably they have made these

[40] See, for example, Gilbert (1996a). Roth explicitly states that the feature of shared agency he is trying to understand is the same one that Margaret Gilbert focuses on (Roth 2004, p. 363).

commitments to each to each other when and by forming the shared intention to walk together.

These *contralateral commitments* explain the entitlements that arise in shared agency, such as Lisa's entitlement to rebuke Zac for walking too quickly. In shared agency, all participants are committed to pursuing the goal of the shared act – this is the participatory commitment. When one adopts a participatory commitment one thereby commits *to oneself* to do one's part in the activity – this is the ipsilateral commitment.[41] A *contralateral commitment* is in many ways to shared agency what the ipsilateral commitment is to individual agency. When you are a participant in a shared action, you make a commitment to the other participants to do your fair share. This corresponds to the commitment you make to yourself in singular agency. And so the entitlements constitutive of shared agency are structurally similar to the entitlements of individual agency.

This idea will be important to what follows, so it is worth pausing and restating again. *Contralateral commitments are to shared agency what ipsilateral commitments are to individual agency.* If you accept the later, you ought to accept the former.

Roth's project is to explain how contralateral commitments come into being. His answer is that they are the product of an individual acting on the intention of another. When A acts on B's intention, A in effect makes a commitment to B to do what is

[41] Ipsilateral commitments, like contralateral commitments, also come with entitlements. One is entitled to rebuke oneself if one fails to do one's part.

necessary to fulfill the intention, and B, in issuing the intention, makes a commitment to A to put A in a position where the intention can be satisfied.

Is the thesis of **Practical Intimacy** an adequate account of shared agency? I will set this question aside for another time. The lessons I want us to learn from Roth's work are that (a) Gilbert is not unique in thinking that shared agency is constituted in part by a normative structure consisting of obligations and entitlements, and (b) this normative structure is structurally parallel to the normative structure of individual agency

With this in mind, let us now turn to Bratman's Planning Theory.

7 Bratman's Planning Theory

7.1 The Details

Bratman approaches the issue of shared agency from a methodological perspective that we can summarize with three broad theses:

(1) Continuity Thesis: "the conceptual, metaphysical, and normative structures central to [shared agency] are … continuous with the structures of individual planning agency" (2014, p. 8).

(2) Individualism: Shared agency can be constructed out of the building blocks of individual agency.

(3) Modesty Thesis: The correct account of shared agency should lie between a minimalist view according to which we can construct shared agency out of nothing more than two individuals acting in accordance with Nash equilibrium, and a robust view according to which agency is constituted by "distinctive interpersonal obligations and entitlements" (2014, p. 11).[42]

[42] Here, of course, Bratman is referring to views like Margaret Gilbert's.

What is this "individual planning agency" out of which Bratman hopes to construct shared agency? The Planning Theory of Agency is the view that intentions are plans and that practical reasoning is, roughly, the process of planning. Forming an intention is setting a plan, or as he sometimes puts it, "committing to a plan of action" (e.g., Bratman 2013, p. 51). The function of planning is to organize our agency over time. This process of planning comes with a package of norms that we sometimes associate with intention rationality:

(a) **Consistency**: our intentions ought to be internally consistent as well as consistent with our beliefs.

(b) **Agglomeration**: we should be able to agglomerate our intentions into one larger intentions that satisfies the norm of consistency above.

(c) Means-end coherence: Our intentions place demands on us to take the required means to the ends contained in those intentions. This often requires making subplans that in turn must satisfy all the norms associated with intention rationality.

(d) Stability: Our intentions, though revisable, are *resistant* to revision.

Planning agency, then, is what Bratman calls the kind of agency associated with intentions and practical reason. The basic idea behind his "planning theory", is that by recognizing that intentions function as plans, we can explain where the norms associated with intention rationality come from. Acting in accordance with these norms is required to make and follow through with plans.

Let us now turn to Bratman's proposal for constructing shared agency out of the components of individual planning agency. Bratman starts this process by reflecting on what the function of a shared intention would be for a planning agent. The most obvious reason that we share our agency with others is to achieve ends that we cannot otherwise

achieve alone. With this in mind, Bratman identifies three functions that shared intentions would serve for planning agents (1999b p. 112):

(1) By sharing our intention with those whom we participate in shared activities, we will be better able to coordinate our respective actions.

(2) Our shared intention will help us coordinate our respective sub plans.

(3) Our shared intention will provide a framework for bargaining about who should do what in the pursuit of our end.

Our account of shared agency should be able to show how shared agency serves these functions.

Bratman's proposes to construct shared agency out of the constituents of planning agency by way of interlocking intentions. The basic idea is that if you and I each intend that we J, and our intentions are interlocking such that we are mutually responsive and supportive of the other's intentions, what we will arrive at is a structure that can serve the functional role that shared intentions can serve. Bratman's account is a bit complicated. Looks take it piece by piece.

Step 1

You and I together intend to J if:

(i) I intend that we J and you intend that we J.

There are two things to note about this step. First, the intention is an intention *that* rather than an intention *to*. Bratman constructs shared intentions out of intentions *that* because they are more permissive in terms of what their contents can be. For example, I cannot intend you *to* do anything – only you have this power – but I can intend *that* you do something, by, for example, ordering you to do it. Second, the content of this intention

includes ‘we’ and the action I intend that we perform. This is not circular because ‘we J’ is meant to be read as “neutral with respect to shared intentionality” (2014, 46). Imagine that ‘J’ stands for “climbing the mountain”. I could intend that we both climb the mountain *without* intending that we climb the mountain *together* as a shared activity. (I could intend that you climb it on one side and me on the other.)

Step 2

You and I share an intention to J if:

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J.

This condition requires that each of us intend that the intention of the other play a role in the joint activity. Borrowing an example from Bratman, imagine that I intend that we go to New York City. But I intend to kidnap you and drive you to New York. While this would (in a certain sense, at least) be a case where we both go to New York, it would not be case where we go to New York as a result of a shared intention to do so.

Step 3

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor of J-ing that mesh with each other.

In order for us to execute our intention that we J, each of us will need to take certain means to J-ing. Imagine that you and I are building a shed in the backyard together. This is a complex task that requires quite a bit of coordination between us. For example, we have to figure out who will work on which side. (We don't want to end up with two fronts, after all!) We will also need to coordinate who will pick up what supplies, and who will bring which tools. These are our subplans, or the means that we intend to take to our ends. To make sure that what we together do turns out to be sufficient to build the shed, we need our subplans to "mesh" – my subplans need to be responsive to your subplans, and your subplans need to be responsive to my subplans. If we both bring nails but neither of us brings wood, we're not going to build much of anything. Our subplans mesh when they are "mutually responsive", that is, when my subplans are responsive to your subplans and your subplans are responsive to my subplans.

Step 4

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor of J-ing that mesh with each other.

(iv) We each believe the following: if each of us continues to intend that we J, then we will J by way of those intentions.

Notice that condition (iv) introduces a belief into our cluster of attitudes. The content of this belief concerns two things: first that our intentions persist, and second that the

persistence of our intentions lead us to take the means necessary and sufficient to our J-ing. If I have this belief, then I must also believe that you will be responsive to my subplans, and if you have this belief you must believe that my intentions will be responsive to your subplans.

Step 5

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor of J-ing that mesh with each other.

(iv) We each believe the following: if each of us continues to intend that we J, then we will J by way of those intentions.

(v) We each believe that our intentions in (i) are persistence interdependent.

(vi) Our intentions in (i) are persistence interdependent

Persistence interdependence is the idea that each of our intentions that we J persists so long as the other person's intention persists. If I believe this about your intentions, and my intentions persist, then I will believe that your intention to J will also persist. If our intentions are persistence interdependent, then all I need to do to make your intentions persist is to maintain my intention.

We need a condition about persistence interdependence because we have no control over each other's intentions. If our shared action requires that each of our intentions persist, this will put me in a precarious position. If you change your intentions,

then my intention that we J will not be efficacious. This means that if I suspect that you might change your intention, it would be irrational for me to keep mine, and the same will go for you. But if I believe that your intention will persist so long as my intention will persist, then (assuming my belief is true) I in effect have control over whether your intention persists: I can simply maintain my intention. Mutual belief about the persistence interdependence of our respective intentions makes our shared intentions stable, which is necessary for them to play their functional role.

Step 6

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor J-ing that mesh with each other.

(iv) We each believe the following: if each of us continues to intend that we J, then we will J by way of those intentions.

(v) We each believe that our intentions in (i) are persistence interdependent.

(vi) our intentions in (i) are persistence interdependent

(vii) Common knowledge of conditions (i)-(vii) exists among us.

You and I have common knowledge about p if and only if (a) I know that p and you know that p, (b) I know that you know that p, and you know that I know that p, (c) and so on. A common knowledge condition is common in accounts of shared agency because

common knowledge is a way in which we can share (a small subset of) our minds. If our minds are locked off from each other, then we cannot come together to share an intention.

Step 7

(i) I intend that we J and you intend that we J.

(ii) (a) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor J-ing that mesh with each other.

(iv) We each believe the following: if each of us continues to intend that we J, then we will J by way of those intentions.

(v) We each believe that our intentions in (i) are persistence interdependent.

(vi) our intentions in (i) are persistence interdependent

(vii) Common knowledge of conditions (i)-(vii) exists among us.

(viii) The connection between the shared intention and joint action involves public mutual responsiveness.

The purpose of condition (viii) is to secure an appropriate connection between intention and action. Condition (ii) and (iv) respectively concern *intentions* and *beliefs* that an appropriate connection will obtain. Condition (viii) states that such a connection *does* obtain. This is required to get from a shared intention to a shared action.

The argument of section 6 made heavy use of the normative components of Gilbert's Plural Subject Theory. As I mentioned in that section, Bratman rejects the idea that mutual obligations are a constitutive feature of shared agency. Now that we have the

details of Bratman's view on the table, we are in a position to consider his arguments. I turn to these next.

7.2 Bratman's Objection to Gilbert

Gilbert argues that an adequate account of shared intention must make sense of the mutual obligations constitutive of shared agency. As we will see, Bratman argues that while many cases of shared intention in fact come with mutual obligations, they are not constitutive of it. There can be shared intentions with no mutual obligations. Bratman defends this claim by way of a case-based argument. I will argue that Bratman's argument against Gilbert gets him stuck in a similar bind to the one that Gilbert found herself stuck in. Let's consider Bratman's first case:

> **CASE 1** Suppose that I tell you that unless you join with me in a shared intention to sing the duet I am going to blow up your house. I thereby coerce you into satisfying your side of [the conditions for forming a shared intention]. But given that your participation is coerced in this way, it seems to me that in this case I have no entitlement to your playing your part.[43] (1999c, pp. 132-133)

From my discussion of the Plural Subject Theory, we can infer that Gilbert would hold that the coercer and the victim in **CASE 1** may in fact have formed a joint commitment, and that, in virtue of this fact, the victim would be obligated to the coercer. I argued that because Gilbert was mistaken in the second of these claims (that the victim would be obligated to the coercer), that she was also mistaken in the first (that this is a case of shared agency). Bratman, on the other hand, wants to say that the first of these

[43] Here Bratman drops the following important footnote: "I do not say that coercion always blocks obligations, only that it can and that it plausibly does in the present case."

claims (creation of a shared intention) is true, but the second claim (the victim is obligated to the coercer) is not. If Bratman is right, this would spell trouble for the Shared Agency Hypothesis. Before providing my response to **CASE 1**, allow me to present Bratman's second case.

> **CASE 2** Suppose you and I each announce our intention in favor of our duet singing, given that the other has the relevant intentions. Indeed, we announce intentions of just the sorts required by the [Planning Theory of Shared Agency]. And you and I each expect our announcement to lead the other justifiably to the corresponding belief. But each of us adds to our announcement the qualification: It is very likely that I will continue so to intend. But I reserve the right to change my mind at will and I recognize that you do too. Neither of us is obligated to the other to continue so to intend. (ibid p. 133)

According to Bratman, in **CASE 2** the two parties have formed a shared intention. But because they have both disavowed that the other has any obligation to satisfy the intention, no such obligation exists, for both parties would lack an entitlement to rebuke the other for not fulfilling their end of the intention. As Bratman points out, Gilbert is committed to holding that either there is no shared intention in this case (because no obligation) or there are in fact obligations in this case, despite appearances to the contrary. Neither of these options seem plausible to him. So he takes this as another strike against Plural Subject Theory.

I think that Bratman is wrong about these cases. I will argue that in **CASE 2** if there is genuinely a shared intention between the two singers, there must be mutual obligations. It is incoherent to suggest otherwise. Given what I argue in defense of this case, I will then argue that **CASE 1** must not in fact be a case of shared intention. Here I will rely on the arguments I provided in previous sections of this chapter. As we will see, Bratman goes wrong because he misconstrues the nature of the obligations in question.

Once we recognize that Gilbert's mutual obligations are of the same normative kind as the so-called rational requirements, Bratman himself will be committed to embracing them.

For the sake of argument, let's assume that Bratman is right about **CASE 2**, that singer 1 has no obligations to singer 2 and singer 2 has no obligations to singer 1. What obligation, specifically, does Bratman think neither singer has to the other? According to Bratman, both of the singers agree to the following: "Neither of us is obligated to the other to continue so to intend." This, of course, is not the content of the obligations that Gilbert (or, as we saw, Roth) identifies as being constitutive of shared intentions. For Gilbert, parties to a shared intention are obligated to each other *to do their part* in pursuit of the end contained in the shared intention.[44] But presumably – at least for in individualist like Bratman – if one disavows the other's obligation to maintain a participatory intention, one also disavows the other's obligation to do their part in contributing toward the shared intention. Whatever the case, I will argue that neither of these ways of interpreting the situation can work.

There are two normative dimensions of intention that will be relevant to the argument I am about to make, both of which Bratman accepts: the stability and means-end coherence of intentions. Recall that the stability of intentions concerns their "stickiness" – though intentions are revisable, there is normative pressure against our

[44] Gilbert doesn't hold that there is an obligation to maintain a personal intention when one is jointly committed. First, she doesn't think continuing to intend is necessary for a shared intention to persist. Second, as we saw in **Olive's Report**, Gilbert doesn't think one has failed in one's obligations *vis-à-vis* a shared intention by giving up on one's personal intention. Olive no longer intended to climb the mountain. But since she continued on up the mountain, she wasn't failing in her obligations to her hiking partner

revising them. In terms of the requirements of rationality, we might say that revising one's intentions requires that one have (*pro tanto*) reason to do so. Means-end coherency concerns making subplans to take the means toward one's end. Put in terms of rationality, we ought either to give up our intentions, or take the means toward satisfying them.[45]

Now Bratman thinks (correctly, in my view), that shared intentions have the same normative dimensions that personal intentions have. So shared intentions must exhibit stability and means-end coherence (among others). Earlier I said that Gilbert thinks that if one makes a *commitment* to do something then absent reason saying otherwise, one ought to follow through on the commitment, and that she thinks this 'ought' "is a matter of what might be referred to as a rational requirement" (2009, p. 179). Notice that what Gilbert is pointing to here is the same phenomenon that Bratman is pointing to: the stability of intentions.

I argued above that for Gilbert, the obligations and entitlements constitutive of joint commitment are in the same normative family what we commonly refer to as the rational requirements pertaining to personal commitments (or individual intentions). In this case, the obligations we have to the other parties in a joint commitment are the analogue of the requirement to take the means to an end. Gilbert says that I am obligated to the other parties in the sense that I am *answerable* to them; they have standing to rebuke me. But this is not different from what happens when I fail to take the means to an end—I have standing to rebuke myself.

[45] This is one way of formulating the instrumental principle. This is to be read – as in chapter 2 – as a wide scope principle.

In light of this, let's consider **CASE 2**. Bratman claims that the singers share an intention but that they are not obligated to each other. This means that neither is in a position to rebuke the other for failing to do their part. This is problematic for two reasons. First, it violates one of his own conditions – *persistence interdependence.* Recall that our intentions are persistence interdependent if and only if your intentions persist if my intentions persist, and vice versa. According to condition (v), each of the parties to a shared intention must believe that the personal intentions of the other party are persistence interdependent, and according to condition (vi) they must *in fact be* persistence interdependent. The relevant intentions of the singers in **CASE 2** are not persistence interdependent because either singer may decide – regardless of what the other intends to do – to rescind their intention. Further, both signers recognize that this is the case. So both of condition (v) and (vi) are violated, so there is no shared intention.

But there is, I think, an even deeper worry. Consider the following sort of case for an individual intention: a person intends a certain end but says to himself "there is no normative pressure for me to take the means to this end. I probably will, but it's not that there is any sense in which I should or ought to." Either this person is being disingenuous, or he is not really setting himself an end. When you set an end, you are under normative pressure to take the means. This is just part of what it is to set an end. If you fail to try to take the means to your end, you are answerable to yourself; you are entitled to be disappointed and rebuke yourself for your failure. But the same goes for a shared intention. If two parties share an intention to E, they commit to taking the means to E. But this is a commitment they make not only to themselves, but also to each other.

Were the two parties to each say that the other was not in fact committed to doing their part in the shared activity since they themselves were not, this would be analogous to not committing to taking the means to an end for a singular intention. "Let's build a shed. But you're not committed to doing your part and I'm not committed to doing my part. But we might still build a shed." This is not sharing an intention to build a shed.

Is **CASE 2** a case of shared intention? My sense is that on either Bratman or Gilbert's view the answer is no. It's not a shared intention on Bratman's View because the persistence interdependence condition is not met. It is not a shared intention on Gilbert's view because neither party seems to express a genuine readiness to be jointly committed. (They don't express such a readiness because both parties too easily disavow the obligations that come with so committing oneself.)

Let's now turn to **CASE 1**. Recall that in this case, one singer threatens the other singer to join her in a duet. Bratman's conclusion about this case is that we have a case of shared intention with no mutual obligation. We've already seen Gilbert's line on this sort of case: while there is certainly no *moral obligation* owed by the threatened singer, Gilbert insists that there *is* still a non-moral obligation the singer owes to the coercer. In section 6, I argued that Gilbert is mistaken in this. Obligations *qua* rational requirements are grounded in autonomous commitments; when someone is coerced down a certain path, no such commitment is made.

What can I say in response to Bratman? His reading of the case is that there are no mutual obligations, but there is still a shared intention. So he seems to agree with me, but draws a different lesson. I have two strategies for responding. The first strategy is to

argue that obligations *do* come with intentions, and then appeal to the argument I offered above which shows that one cannot unilaterally impose such obligations on another. One might worry, however, that this strategy is question begging for it assumes the very thing that the case is meant to reject – namely that there cannot be intentions without mutual obligations. Though I don't think this is question begging – we have good reason to think that such obligations must be a constitutive aspect of shared intentions – there is a second way we can address CASE 1. The two parties in the case do not share an intention because conditions (ii) and (iii) of Bratman's "Planning Theory" are not satisfied.

Consider first, condition (ii):

> (ii) I intend that we J by way of my intention that we J and your intention that we J; (b) you intend that we J by way of your intention that we J and my intention that we J

I argued in the previous section that when you coerce me, you force me into a normative landscape with only one path forward: do as you demand. When you demand that I sing with you, what is the content of my corresponding intention? It is to do as you say, to avoid the harm that you are threatening me with. My intention is certainly not for us to sing by way of your intention that we sing and my intention that we sing. It is that *I* will sing. And I will be singing because you are forcing me to. But I certainly don't have the intention that *you* sing by way of your intention that we sing. I could care less about what you do, and in fact hope that you drop dead right now so I can stop being your puppet.[46]

[46] Perhaps you threaten me in this way: If we don't harmonize, then I will kill you. Now we might think that you have forced me to intend that you sing by way of your intention that we sing. But I don't think this is right. I now intend to harmonize with you and hope that you do the same, but frankly I don't care how it happens, as long as you don't kill me.

Now consider condition (iii):

(iii) We each intend the following: that we J by way of sub-plans of each of our intentions in favor of J-ing that mesh with each other.

The argument here will be similar to the one just given. When you threaten me, my intention is to do as you tell me to do, which in this case is to sing with you. But in doing so, I do not intend that you take the appropriate means to achieve your end. My concern is avoiding your violence, not the efficacy of your agency vis-à-vis our singing a duet.

But that coercion is inconsistent with Bratmanian shared intentions does not mean that Bratmanian shared intentions require both parties to comply with the moral law, as I am understanding it. Is there a general argument that we can offer to show that, in fact, Bratmanian shared intention does in fact require all parties to the intention to comply with the moral law? Here the argument I would offer is the one advanced above when addressing Gilbert's view. Shared agency has a normative structure which cannot be imposed on another unilaterally. As I've argued, Bratman must accept this no less than Gilbert. In chapter 2, I argued that one must commit to an enterprise in order for the enterprise's normative structure to be binding on one. But when one is used as a mere means, one does not commit to the enterprise that the other party is imposing on one. When you force or deceive me into going on a walk with you, you are trying to force me into a shared activity. But because I am forced or deceived, I do not commit to a shared activity, so understood. *You are dragging me into an enterprise that I did not commit to*.

To exercise agency is to commit to doing one's part in pursuing an end. When I act alone, I make this commitment to myself. When I act with another, I make this commitment to them as well as to myself. Having made this commitment to you, and you

having made the same to me, we are now entitled to each other's participation. This is nothing more than the instrumental principle, distributed across persons.

8. Conclusion

This was a rather long and complicated chapter. Allow me to briefly tie all the pieces together.

In part I, I argued that Korsgaard's Collective Action Argument fails because it equivocates on the concept of public reason. The point of part II was to argue that the conclusion of this argument, which I have been calling the Shared Agency Hypothesis, can nevertheless be defended. Here is the argument in a nutshell. Shared agency has a normative structure, much like singular agency. This normative structure is written into the nature of shared agency itself. To engage in the activity of shared agency is to be bound by the norms constitutive of that activity. But to be bound by the norms of shared agency I must commit myself to that activity, by way of an autonomous expression of my will. When you use me as a mere means, you are narrowing the normative space in which I can express my will to a pinhole. The choice I make in such a situation is not an expression of my will, but rather an expression of your will. And as an expression of *your* will, the norms constitutive of the activity into which you are trying to force me do not bind me.

If the preceding argument is correct, we have discovered something rather surprising, both about our agency and morality. Morality grows out of the constitutive

features of shared agency. But how does this help us secure an objective morality? I turn to this question in our final chapter.

Chapter 5: Conclusion – Toward a Social Constitutivism

1. Introduction

My aim in this dissertation has been to develop a version of constitutivism that takes the constitutive principle strategy seriously. I have been arguing up to this point that when coupled with the Commitment View of practical reasoning, we can take it quite far. One thing that I haven't yet shown is how this commitment based-account of constitutivism can capture morality's purported objectivity. Unfortunately, I am not yet in a position to fully answer this question. In its stead I will offer a promissory note of what the form of an answer must look like if we are to succeed on this front, together with some suggestions on how we might fill out a fuller answer. The way to read this chapter, then, is not as an argument that the commitment-based account I've developed works better than competing theories, but rather as roadmap for where those friendly to the view should begin looking.

2. Where We are, How We Got Here, and Where We Need to Go

In the introduction of this dissertation, I stated that this project was something of an experiment. How far could we take an anti-realist view that grounds morality in constitutive features of our agency? Is there a workable version of constitutivism that, unlike the constitutive aim views, somehow makes the moral law a constitutive principle of agency? The view I have been developing in this dissertation aims to do just this. In chapter 4 I defended what I called the Shared Agency Hypothesis, according to which the

moral law is a constitutive principle of shared agency. What this means is that, in order for two or more people to share their agency with each other, they must not treat each other as mere means. The Shared Agency Hypothesis satisfies one of our desiderata: it identifies the moral law as a constitutive principle of agency, albeit *shared agency.* It also helps us solve the Bad Action Problem, the focus of chapter 3. According to the view I develop, there is plenty of room for bad action: these are actions where we treat others as mere means. It follows from the Shared Agency Hypothesis, however, that we do not share our agency with those that we treat in this way.

Of course, establishing that the moral law is a constitutive principle of a certain activity does little, by itself, to establish that the principle is normative. As chapter 2 argued, to show that it is normative, we must show that we are committed to the activity that the principle is constitutive of. The last step in our argument, then, must involve showing that we are committed to engaging in the activity of shared agency. Showing that we are so committed will show that the principle is normative.

But now, here comes the rub. If the moral law is normative only if we are committed to sharing our agency, then this will mean that we've been impaled by the second horn of Katsafanas' dilemma: the moral law will be optional. We will be able to escape the force of the moral law by giving up our commitment to shared agency. If we wish to capture the universality and categoricity of morality, what we will need to show, then, is that we have an *inescapable* commitment to sharing our agency, that it is somehow something that we carry with us no matter what we do. The first place we shall look for an answer is Korsgaard herself.

3. How Korsgaard Captures Morality's Categoricity

In part I of chapter 4 we looked at an argument of Korsgaard's that I called the Collective Action Argument:

1. Collective action requires shared deliberation.
2. Shared deliberation requires acting on public reason.
3. Acting on public reason requires complying with the moral law.
4. So, collective action requires complying with the moral law.

The aim of this argument is to show that the moral law is a constitutive principle of shared agency. Ultimately, we saw that the argument is not sound because it equivocated on two different senses of "public reason". I want to revisit this argument, not because I think we can save the concept of public reason, but because Korsgaard adds additional premises to it. This will be our jumping off point for reflecting on how we can capture the categoricity of morality in the Commitment View.

After defending the Collective Action Argument, Korsgaard writes the following:

> Personal interaction … is quite literally acting with others. But for creatures who must constitute our identities, it is equally true that acting is quite literally interacting with yourself. (2009, p. 202)

If we affix this line of thought to the collective action argument, we get the following:

1. Collective action requires shared deliberation.
2. Shared deliberation requires acting on public reason.
3. Acting on public reason requires complying with the moral law.
4. So, collective action requires complying with the moral law.
5. But singular action is performing a collective action with yourself.
6. So, singular action requires complying with the moral law.

This argument adds only one more premise to the collective action argument before generating a new conclusion. This is premise (5): *singular action is performing a collective action with yourself*. If performing collective action requires complying with

the moral law, and singular action is, itself, collective action, then all action requires complying with the moral law! Let's call this new argument *Korsgaard's Master Argument*.

Before we discuss how Korsgaard defends premise 5 we should take note of an obvious problem this argument faces: *the bad action problem*. Premise 6 appears to make the very move that we went to great lengths to try to avoid in chapter 3: if complying with the moral law is a requirement for singular action, then every action will be good, and there can be no bad actions. This in itself gives us reason to reject Korsgaard's Master Argument as a non-starter. But let's, nevertheless, set this damning worry aside and consider how Korsgaard defends premise 5.[1]

3.1 Korsgaard's Defense of Premise 5

Korsgaard's defense of premise 5 is largely case drive. The argument involves reflecting on one case in particular, which she borrows from Derek Parfit (1984): the Russian Nobleman Case. Here is the case, as she describes it, in full:

> The nineteenth-century Russian is now, in his youth, a socialist, and he plans to distribute large portions of his inheritance, later, when he comes into it, to the peasants. But he also anticipates that his attitudes will become more conservative as he grows older, and that he may not think that this is the right thing to do when the inheritance is finally his own. So he makes a contract *now*, to distribute the land when he gets it, which can only be revoked with the consent of his wife, and he asks his wife to promise not to revoke it then, even if he tells her then that he has changed his mind, and that she is released from the promise. (2009, p.185)

[1] Perhaps the bad action problem is not an inevitable consequence of this argument. One solution is to distinguish synchronic from diachronic action. Perhaps the moral law is a constitutive principle of diachronic action but not synchronic action. This has the unfortunate implication that there are no bad diachronic actions, which seems incorrect.

Korsgaard goes on to add that

> The young Russian does not anticipate that he is going to become irrational, that his judgment will be clouded, or that the immediate temptation of having the estates will undermine his self-control. He simply believes that when he is older he is going to have different values than the ones he has now. (ibid)

The relevance of the Russian Nobleman Case to the defense of premise 5 is that, as Korsgaard describes it, the Russian nobleman is a paradigmatic case of a person who is disunified. According to the young nobleman, the old nobleman is akin to an imposter: the wife is not to listen to him, not to give the old nobleman "a voice in the disposition of the estates" (186). Korsgaard quotes Parfit as saying "It might seem to her as if she has obligations to two different people" (p. 185; qtd. in Parfit, p. 327).

There are two questions we can ask about this case. First, what should the wife do? How should she we treat her husband? This is an interesting question – and certainly a philosophical puzzle in its own right – but it's not the question that concerns Korsgaard. The question that concerns her is *what should the Russian nobleman do*?

To answer this question, we have to first diagnose what the Russian nobleman did *wrong*. According to Korsgaard, the young Russian nobleman isn't treating his future self with respect. He's not treating his future self with respect in the same way that I would not be treating you with respect if I asked someone to coerce you into performing an action that was contrary to your better judgment. According to Korsgaard, the Russian nobleman treating himself this way is the source of his disunification.

From this case Korsgaard produces the following argument:

> The requirements for unifying your agency internally are the same as the requirements for unifying your agency with that of others. Constituting your own agency is a matter of choosing only those reasons you can share with yourself.

> That's why you have to will universally, because the reason you act on now, the law you make for yourself now, must be one you can will to act on again later, come what may, unless you come to see that there's a good reason to change it. (2009, p. 202)

The Russian nobleman goes wrong, then, because he fails to act on a reason that his future self can share. He is treating his future self as a person to be sparred with, rather than a person to *interact* with.

I think that Korsgaard is right that the Russian nobleman is lacking self-respect. It seems that one of the central tenets of self-respect is trusting yourself. Just as it is wrong to force someone into a certain course of action if they disagree with your line of reasoning, it is wrong to box yourself into a course of action simply because you might come to change your mind later on.[2] The correct attitude to take toward yourself is the same that you should take toward others; you should trust in your ability to make normative judgments. When someone disagrees with us, we should try to reason with them and convince them of our point of view. But we should also try to understand where they are coming from – perhaps *we* are the ones who are wrong. The Russian nobleman should also trust himself. If he genuinely thinks that he will change his mind, and not as a result of weakness-of-will or irrationality, then he should reason through the case and try to determine why he will change his mind. If he is still unsure, perhaps he can write down a plea to his future self, reminding him why he should give the estate away.

[2] The second clause of this sentence, if it is to be true, must include an implicit *all else being equal* clause. Many of the more important choices we make in life – choices concerning our career, our family, and the pursuit of what we deem of value – result in boxing ourselves in. For example, I might now decide to have a kid. But this choice will result in precluding certain courses of action that I might otherwise pursue were I not to have a kid. The same is true for certain career paths we might pursue. Were I to decide to hang up my guitar and pursue academia, this would preclude me from a career in music.

But does it follow from the fact that the Russian nobleman is lacking self-respect that singular action is collective action with oneself?

4. Problems with Korsgaard's Argument

According to Korsgaard, we must act on the moral law in order to unify our diachronic agency. Singular action is collective action because unifying my diachronic agency requires sharing my agency with my future self. There is a problem with the underlying idea here. I can't share my agency with my future self, at least not in the sense that we mean when we are talking about collective action. Consider Margaret Gilbert's Plural Subject Theory. According to Gilbert, two parties enter into a joint commitment when they each express a readiness to do so under conditions of common knowledge. But I can't do this with my future self, for my future self can't express his readiness to me, as he doesn't yet exist. Further, when he does exist, I will no longer be around. If neither of us is around when the other is in a position to express his readiness, our mutual readiness cannot be expressed in conditions of common knowledge.

Since Michael Bratman's "Planning Theory" also has a common knowledge condition on shared agency, a similar argument shows that diachronic singular agency cannot be a form of shared agency on his account of shared agency either. But there is an additional argument to be made in this case. Recall that Bratman's "Planning Theory" includes a condition according to which the intentions of the parties to the shared intention must be *persistence interdependent*. The intentions of person A and person B are persistence interdependent if and only if A's intention persists if B's does, and B's

intention persists if A's does. Korsgaard's idea of diachronic agency as shared agency violates this condition because it would require that backward causation be possible (a relationship between B retaining his intention and A retaining his).

Of course, these arguments don't say anything about the question of whether the moral law is a unifying principle of diachronic agency. The point, rather, is that arguing for this conclusion by way of the collective action argument just can't work.

There is another, perhaps deeper, though less technical reason for rejecting the move that Korsgaard makes here. The moral law *qua* constitutive principle of shared agency is, I think, an attractive idea. It gets, what we might call the "directionality" of morality correct. Being moral is about interacting with *others*. The deeper worry about Korsgaard's attempt to capture the categoricity of morality is that it does so at the expense of capturing this important aspect of morality. While the Collective Action Argument, as we interrogated it in the previous chapter, at least on its face appropriately captured morality's other regarding nature, the move inward toward satisfying the conditions of singular diachronic agency threaten to completely swamp that out. In fact, if we reflect on how Korsgaard's Master Argument works, we can see that collective action need play no essential role in her view. What *really* matters for capturing the universality and categoricity of morality is that the moral law is a constitutive principle of *diachronic singular action*. As far as the moral law is concerned, there need be no other people to interact with– we could all be islands unto ourselves. Capturing morality's categoricity by making conformity with the moral law a solipsistic enterprise is, I think, the wrong move for constitutivism. We shouldn't sacrifice morality's directionality in

order to capture it's categoricity. In what follows I will sketch a proposal for developing the commitment view in a way that can perhaps capture morality's categoricity while at the same time securing it's other regarding nature. The basic idea is that we should understand our agency as essentially social in nature. To put this in terms of the Commitment View: we have an inescapable commitment to sharing our agency with others.

5. Going Social: A New Wave?

In the past couple of years there seems to be a small wave of philosophers taking constitutivism in a social direction. The insight that these philosophers are having is similar to the one that I am latching onto in this dissertation. Our nature as social beings should play a role in our foundational accounts of morality. In this section I discuss in brief two recent contributions to this new wave: Kate Manne's practice-based account of normative reasons, and Kenneth Walden's social constitutivism. I discuss the later of these in more detail in Appendix A.

5.1 Kate Manne's Practice-Based Account of Normative Reasons

In a recent article, Kate Manne (2013), articulates an account of a certain domain of practical reasons, which she calls the "practice-based view" (p. 51). According to this view, a "certain type of reason" is grounded in "facts about *what we do*, or about *what one does*, as a participant in certain sorts of collective practices, joint enterprises, or particular social relationships" (pp. 51-52).

The practices that Manne has in mind are those social practices which are "constituted" by norms in such a way that the activities "cannot be characterized independently of [the norms]" (p. 54). Manne quite clearly has in mind constitutive principles. In fact, the example she appeals to when eliciting reader's intuitions is none other than the familiar baseball yarn:

> [P]art of what makes baseball the game it is is the 'three strikes and you're out' rule. This rule could potentially be changed, but it would make baseball a somewhat different game… Moreover, baseball as an activity cannot be characterized independently of its rules, whatever they happen to be at the time. (p. 54)

Now, Manne thinks that there are many (non-game-based) social practices that are also constituted by norms. Her primary example is friendship:

> Friendship, for instance, would not be what it is (I propose) without its characteristic norms, such as loyalty and trust. A disloyal and distrustful friend is, in the first instance, a bad friend—and eventually, if the disloyalty and distrust persists, they are not really a friend at all. (ibid)

In this passage, Manne is suggesting that loyalty and trust are constitutive norms of friendship. If you don't treat an individual with loyalty and trust, you are not treating them as a friend.

Manne's practice-based view is quite friendly to Social Constitutivism. In this passage we find that she appears to think that there are certain norms that one must conform to in order to engage in the practices that those norms are constitutive of. The

friendliness doesn't stop here though. Manne wants to argue that "social practices themselves provide … reasons" (pp. 63).[3] She adds that,

> On a practiced based view, *all* reasons thus generated have a partial cast—that is, they are generated by local facts about one's particular commitments, and what these in turn require one to do as part and parcel of the relevant social role. (pp. 63-64)

So for Manne, practice-based reasons are generated by an individual's commitment to engage in a practice. This strategy for generating normativity out of descriptive features of the world is similar to the commitment strategy I descibe in chapter 2 of this dissertation.

There is an important dissimilarity between our views that is worth noting before moving on. Manne's view rests the genuine normativity of practice-based reasons on the back of a broadly consequentialist moral theory. We *really* only have reason to do what is constitutive of the social practices to which we are committed if these social practices provide some general utility to society (friendships do, mafia membership doesn't). On what these consequentialist considerations rest is not clear to me.

5.2 Kenneth Walden's Social Constitutivism

In a recent article, Kenneth Walden (2018) – one of the few philosophers who recognize that Chapter 9 of *Self-Constitution* "is pivotal for Korsgaard's project" (p. 72) – offers a criticism of Korsgaard's argument that is similar to mine:

[3] What she means by this is that, e.g., the source of my reason to treat my friend with loyalty is the social practice of friendship itself, not a reason external to the nature of the social practice to which my friend and I are together engaged.

> Because [Korsgaard's view does] not include other people as essential constituents of the conditions of agency, [it] cannot construct a constitutivist validation of the claim that we owe something to these other people as such. (p. 79)

Walden proposes a version of constitutivism that he argues does not fall prey to the agential solipsism both he and I find in Korsgaard's view. I won't wade too deep into the complex details here – those interested can turn to Appendix A – but rather will touch on the barebones of his account and briefly discuss the similarities and differences between his view and mine.

The basic gist of his view is that intentions are such that forming them commits us to take into account the normative authority of other agents to make claims on the interpretability of our actions.[4] The authority that we are committed to recognize here is a very thin one and doesn't yet get us into moral territory. But Walden argues that this commitment "work[s] like the thin edge of a wedge." (98) It leads us, by way of the holistic requirements of normative judgements – in particular the standards of consistency and coherence – to the commitment that we recognize the authority that people may have in all the claims that they make (ibid).

From what little I've given the reader, it won't be clear how Walden gets from his starting point to the conclusion. (Again, for this, the reader can turn to the Appendix.) Our concern here is with how Walden's brand of constitutivism – which, like me, he calls social constitutivism – relates to the version of social constitutivism I have been developing in this dissertation.

[4] The reasons for this are complicated, which is why I set them aside in the appendix.

The first point of similarity concerns the fact that Walden does not adopt the constitutive aim strategy; he nowhere attributes a constitutive aim to agency. Is his view an instance of the constitutive principle strategy? Not quite. While he does identify constitutive conditions of intention that one must satisfy in order to count as intending (these are the conditions of interpretability), he doesn't identify a fundamental principle of morality which serves as a constitutive principle of an activity.

The second similarity between our two views concerns the most obvious similarity. We both take our sociality to be a crucial element in making constitutivism work. However, what this sociality *amounts to* is different. For Walden, the social element out of which morality sprouts is the public nature of our intentions. For me, the social element is far more robust: the way in which we share our agency with others.

5.3 Borrowing from My Fellow Travelers

One question worth reflecting on is what we can learn from Manne and Walden to further develop the version of social constitutivism I have been defending in this dissertation.

In Manne we found a philosopher who seems friendly to the idea that a commitment to an activity generates normativity out of what is descriptive of that activity. In Walden, we found a fellow constitutivist bothered by the tendency of mainstream constitutivism to turn morality into a solipsistic affair. When we combine these two philosophers together, we arrive at something that, broadly speaking, approaches my target view.

From Walden, we would benefit by reflecting on how our commitment to sharing our agency with *some*, might "work like the thin edge of a wedge." Might this commitment somehow lead us, by way of coherency and consistency requirements, to a commitment to treating everyone as an end-in-themselves?

At the end of her essay, Manne makes what might be considered a throw-away claim:

> For, we might hold that there are very general social practices—perhaps even social norms—which apply to human beings as such, i.e., *qua* interacting, social creatures, who are the mutually intelligible objects of friendship and love. If we do recognize the existence of such an overarching practice of common humanity, then we might hope to defend the radical claim that all of the moral reasons that pertain to our treatment of our fellows are ultimately practice-based. (pp. 69-70)

This, I think, is what the social constitutivist, at least as I am developing, should go after. Might there be some social practice, similar in relevant ways to shared agency – perhaps the social practice requires sharing our agency – to which we are all inescapably committed?

6. Steps Toward Moral Objectivity for Social Constitutivism

We ended the last section with some vague gestures about how to take social constitutivism to the holy grail of moral objectivity. The question we are asking is this: is there a way to harness the Shared Agency Hypothesis such that we can secure the universality and categoricity of morality?

6.1 Shared Agency is a Practice We are Committed to

The move that the social constitutivist needs to make is argue that sharing our agency is a practice that we are committed to. This, of course, is the move they *have* to make, because, as we saw in Chapter 2, it is in virtue of being an activity to which one is committed that the constitutive principles of that activity become normative for an individual. Here it will be helpful to borrow from Kate Manne's practice-based account discussed in the previous section.

But showing that sharing our agency is an activity to which we are committed is, by itself, insufficient. The Social Constitutivist has to show that the *scope* if this commitment is right. In particular, what has to be shown is that I violate this commitment when I act immorally, even when, in so doing, I did not intend to share my agency with anyone. Further, it must somehow be shown that I am in *some sense* committed to sharing my agency with *everyone*.

What also must be shown is that this commitment is *persistent* – it remains even when we do not, at the time, have a corresponding intention – but also *inescapable*. If we want to show that the moral law is a categorical and universal principle, we must show that the commitment is not *optional*. What must be shown, then, is that our commitment to *sociality*, as I will call it, is akin to our commitments to efficacy and consistency. If we are to show that the moral law is a categorical and universal normative principle, we must show that it's normative authority is grounded in a *constitutive commitment* – a commitment constitutive of the kind of beings that we are.

6.2 Sharing Our Agency is Important

Why think that a commitment to sociality is a constitutive commitment of the kind of beings that we are? The first thing that the social constitutivist should point out is that sharing our agency is important to our way of life. Many of our life projects – being a member of a profession, creating and experiencing art, participating in athletic competitions, forming families and friendships – are inherently social projects in at least two different senses. First, many of these activities involve the participation of other people in such a way that we share our agency with them. Second, these activities take place within a rich social tapestry. In fact, they cannot take place outside of a social framework. Engaging in practices like these are central to the way that we structure and conceptualize our lives. To give up our commitment to sociality would require giving up a core component of who we are, a central aspect of our self-conceptions. I imagine that for most of us, this is something which we simply cannot do.

This idea is quite reminiscent of Korsgaard's concept of a "practical identity". For Korsgaard, a practical identity is a description under which we value ourselves (1996a). Using the notion of a commitment, we could reformulate Korsgaard's notion of a practical identity as so: a practical identity is a social role or social practice to which you are committed. We might think, as Korsgaard seems to think is the case with her formulation, that we all must have *some* practical identity. If this is the case, then there is *some* social role or social practice to which we are all committed.

There are problems facing the social constitutivist who wishes to go in this direction. Even if we accept that there must be *some* social role or social practice to

which we are all committed, we are not yet able to derive from this a commitment to complying with the moral law in general. At most it would appear that we can show that I am committed to treating the other parties in the social practice to which I am committed as ends, but not those outside of my practice. Consider, for example, a mafioso, who is committed to working with his fellow mafia members to run racketeering scams. This mafioso is sharing his agency with his fellow mafia members, and so treats them as ends in themselves, but obviously doesn't treat those he's running a racketeering scam against as ends.

What the social constitutivist must argue, then, is that our commitment to sociality doesn't just commit us to sharing our agency with others, but rather commits us to treating others in such a way that they could share their agency with us. About the mafioso, the social constitutivist must say that his commitment to sociality doesn't just commit him to sharing his agency with the other mafia members, but rather it commits him to being a potential partner in agency even with those against whom he wants to run a racketeering scam.

Why might this be? One potential answer is that we might just never know who we might need to share our agency with in order to achieve our ends. The thought here is that, because we are imperfect agents, and our ends are complex, success in the pursuit of our ends will require sharing our agency with others. But insofar as we can't be sure from whom we will need this help, we are committed to treating everyone as potential parties to a shared action.

The problem with this approach, beyond the dubious idea that everybody is a potential partner in agency, is that it suffers from the same problem that afflicts Korsgaard's Master Argument: we've turned the directionality associated with following the moral law back inward toward the individual agent. We will have captured the morality's categoricity at expense of morality's essential other-regarding nature.

So where does this leave the Social Constitutivist who wishes to show that morality is categorical and universal? My suggestion is that we expand our understanding of what our commitment to sociality involves. Perhaps we should conceive of it, not just as a commitment to share our agency with others, but rather as a commitment to living with others, a core component of which is an openness to sharing our agency with all those willing to share their agency with us.

7. Conclusion

What has this dissertation accomplished? It would seem that I haven't quite achieved the lofty goal I set out to achieve when I started work on this dissertation, namely to show how a social version of constitutivism can provide a foundation for a categorical and universal morality. But even if I failed to do this, my dissertation nevertheless made a few contributions, which I will outline here before closing.

First, my dissertation showed where the source of the normativity of constitutive principles lie, namely in our commitments to engage in the practices that the principles are constitutive of. On my way to showing this, I sketched an account of practical reasoning, the primary constituents of which are commitments. I also argued that the

instrumental principle is a constitutive principle of agency, and that it's normativity is explained by our commitment to acting.

Second (or maybe fourth, depending on how we count), my dissertation showed that a constitutivist who adopts the constitutive principle strategy can nevertheless solve the bad action problem.

Third, and this is perhaps the most surprising of the theses I defended, I argued that the moral law is a constitutive principle of shared agency. If my arguments defending the Shared Agency Hypothesis are successful, this will have been a significant finding, even if, from it, we can't show that morality is categorical or universal. For one, it provides a naturalistically respectable foundation to morality. If the moral law is a constitutive principle of shared agency, this means that morality is part of the natural world. We can expect it to arise whenever beings are able to engage in genuine shared agency. But second, the Shared Agency Hypothesis provides some insight into why we take morality so seriously. On the social constitutivist view, morality is important to us because sharing our agency is important to us.

So while I have only, at best, gestured toward a possible path forward for the social constitutivist to provide a foundation for a categorical and universal morality, along the way we have discovered some interesting and surprising things about the relationship between morality and our agency.

Appendix A: On Kenneth Walden's Social Constitutivism

1 Introduction

In chapter 5, I pointed out that I am not the only social constitutivist on the block. Kenneth Walden (2018) has recently defended a version of constitutivism that is in some ways similar to mine. In fact, like me, he calls his version of constitutivism "Social Constitutivism". The purpose of this appendix is to provide an analysis of Waldens' important contribution to the literature on constitutivism. We will see, however, that Walden's view runs into a similar problem that we ran into in chapter 5.

2. Analysis of Walden's Argument for Social Constitutivism

It is important to keep in mind that in the body of my dissertation, I have been using the phrase "Social Constitutivism" as a proper noun to refer to the view I develop therein. We can represent the view like so:

Social Constitutivism: Commitment View + Shared Agency Hypothesis

If we wish Social Constitutivism to provide a foundation for the Kantian conception of morality described in chapter 1, then the view will also need to include an account of how our sociality is inescapable. We'll come back to this concern at the end of this appendix.

Walden is not a social constitutivist in the sense that he holds the package of views described in this dissertation. Rather, Walden, like me, thinks that morality is grounded in our sociality. Walden, also like me, hopes to show that our sociality is an inescapable feature of our agency, or, as Walden puts it, that "other people [are an]

essential constituent of the conditions of agency" (p. 79). However, I think Walden – again, also like me – doesn't succeed on this score. Showing where his argument fails will be instructive for thinking about the challenges that face the social constitutivist.

Walden's main idea is that the nature of intentional attitudes is such that reflective agents (like ourselves) are committed to respecting the 2nd personal authority of others simply in virtue of exercising our agency. The argument that Walden makes on the road to this conclusion is quite difficult and complicated. As such, I think it will be helpful to start with an outline of it, so we can keep Walden's goal in mind as we weave through each step. The following representation of Walden's argument comes from Walden himself (pp. 93-94):

1. For a creature to be an agent, she must have intentional attitudes.

2. For a creature to possess an intentional attitude, she must be interpretable as having that attitude.

   Lemma 1: Interpretability is a condition on agency.

3. The constitutive requirements of reflective agency have universal normative authority.

4. Reflective agency requires sensitivity to the performance conditions of the agent's prospective actions.

   Lemma 2: Reflective Agency requires sensitivity to the conditions of interpretability.

5. Sensitivity to the conditions of interpretability requires the recognition of the second-personal authority of interpreters to make claims of intelligibility.

6. So, recognition of the second-personal authority of interpreters to make claims of intelligibility is a condition on agency.

These six premises lead to this conclusion:

> (R) It is a universally authoritative requirement that one recognize the second-personal authority of interpreters to make claims of interpretability.

Let's consider the argument in greater detail. In order to make the exegesis easier to comprehend, I've divided the discussion into four parts that correspond to what I consider to be the four major moves that Walden's argument makes.

2.1 Interpretability is a condition on agency.

The first move in Walden's argument is to show that *interpretability* is a condition on agency.[1] This move is fairly straightforward (depending on how we understand it), and, in my view at least, fairly uncontroversial. Following Wittgenstein, Walden argues that intentional attitudes by their very nature must have public content. In virtue of having public content, every intentional attitude – an important instance of which is the attitude of intention – is *in principle* interpretable by a suitably situated interpreter. Because every intention, by their very nature, is interpretable, it follows that interpretability is a condition on agency.

As we will soon see, Walden seems to vacillate between two different senses of interpretability: interpretability as publicity and a much richer notion of interpretability which we will encounter shortly.

[1] Walden's version of constitutivism operates in terms of constitutive *conditions*. While Walden doesn't define what he takes a constitutive condition to be, I take it that a constitutive condition of X is a condition one must satisfy in order to engage in the activity of X-ing. Constitutive conditions, so understood, can be easily translated into constitutive principles. As such, Walden's view is friendly to the constitutive principle strategy I employ in this dissertation.

2.2 Reflective Agency requires sensitivity to the conditions of interpretability.

The next move in Walden's argument is to show that sensitivity to the conditions of interpretation on the part of the agent acting are implicated in the exercise of her agency. To get to this point, Walden first narrows the category of agency he is focusing on to that of *reflective agency*. According to Walden,

> Reflective agency involves reflecting on the reasons one has to perform certain actions and being sensitive to considerations favoring or opposing those actions in a systematic way. (2018, p. 86)

I'm not sure what Walden means by 'involves' here. Is he identifying a constitutive condition? Something weaker? The most common way of spelling out reflective agency is to do so in terms of *capacities*, for example the *capacity* to reflect on one's own attitudes and mental states. Perhaps this is what Walden means by 'involves'. But he might mean something stronger. He could, for example, mean for it to be read in terms of *necessity*: reflective agents *necessarily* are sensitive to the considerations favoring or opposing actions. However, the sentence so understood is surely false. It is quite common for reflective agents to fail at being sensitive to reasons.

Whatever the case, it is certainly true that in *some sense* of 'involves', reflective agency involves sensitivity to considerations favoring or opposing actions. Walden is interested in one subset of such considerations, those concerning the possibility conditions of actions. Walden provides a list of examples of agents who have failed in being sensitive to such conditions:

> If a man attends a baseball game with the intention of umpiring from the bleachers, if her persistently tries to buy things without any money, if he tries to

> knight his beer buddies, if he plans to take my rook in a game of checkers, if he habitually sets about to eat quantities of food greater than his own mass, if he plans to life the moon with a garden trowel, then we would be right to say that something has gone wrong with his reflective agency. For he has failed to apply one important kind of scrutiny in this practical reflection: the regulation of one's actions in light of the conditions on the possibility of their performance. (2018, p. 87)

We find another piece of squishy language in this passage. What does it mean to say that "something has gone wrong with his reflective agency"? Are each of the agents in this list failing to form an intention? Are they failing to perform an action? Or are they acting, just in non-sensical ways?

When it comes to people intending to perform the impossible we can distinguish between two kinds of behavior:

(1) A person who intends to do something *not knowing* that it is impossible to do.

(2) A person who intends to do something*, knowing full well* that it is impossible to do.

The kind of behavior we find in category (1) is not problematic, at least from the standpoint of the exercise of agency. This is nothing more than acting under ignorance. While such a person won't be able to achieve their end, the formation of such an intention is not impossible.[2] Whether the kind of behavior we find in category (2) is possible, however, is debatable. Some theorists in the philosophy of action (e.g., Wallace 2006) think that in order to intend a certain end, one must believe the end to be possible to achieve. Others (e.g., McCann 1997) think that there is no such requirement.

[2] For example, I might intend to ask Robin Williams to sign my VHS copy of *Dead Poets Society*, not knowing that he is dead. This is a perfectly coherent intention to have, even if it is impossible to satisfy.

Unfortunately, Walden doesn't say much about this motley crew of individuals. Are they acting under ignorance or in full knowledge that what they are trying to do cannot be achieved? Perhaps his only claim is that these individuals are acting *oddly*. But if this is all his claim amounts to, it's not clear how he generates from these examples the following lesson:

> [R]eflective agency requires sensitivity to the conditions on the actions that one may entertain performing. Thus, agents must be aware of, and, to some extent, guided by the conditions on their performance of a candidate action. (ibid)

If Walden thinks that individuals who have behaved in the ways listed are not performing actions, that such behavior *qua* action is impossible, then it would seem to follow that reflective agents necessarily must be aware of the possibility conditions on action performance and be guided by those conditions. But if he just thinks that such behavior is "odd", it's not clear how this could support a strong necessity claim about what is "required" for agency. (Odd actions are, after all, still actions.) Perhaps Walden does not mean that agents must be *aware* of *all* conditions, but rather ready to take-up those considerations that *present themselves* to individuals when they are engaged in practical reflection. If this is what Walden means, then his statement is surely true. But, without further detail, it's not clear how the list of examples supports this conclusion. (Again, for all we know, these individuals are simply acting under ignorance.)

We can bracket the second sentence of that passage (the one concerning awareness), however, and accept the first. Surely it is true that reflective agency requires sensitivity to the conditions on the actions that one may entertain performing, at least in some sense of 'sensitivity'. If I am a reflective agent, it would seem that I must be, at the

very least, *disposed* to give weight to evidence that a certain course of action is impossible to perform.

From this idea, Walden derives the following "lemma":

> Lemma 2: Reflective agency requires sensitivity to the conditions of interpretability.

If interpretability is a constitutive condition of an intention, and reflective agency (which involves the formation of intentions) requires being sensitive to the conditions of possibility for acting, then it would seem to follow that reflective agency requires being sensitive to the conditions of interpretability since an intention that is not interpretable is not an intention.

2.3 Sensitivity to the Authority of Interpreters

What conditions of interpretability must agents be sensitive to? On the way to answering this question, Walden points out that interpretation is an institutional enterprise (rather than a set of brute facts).[3] Engagement in an institution requires being "sensitive not only to facts, but also to other people who exercise authority by virtue of the relevant institution" (p. 88). In support of this idea, Walden asks us to consider the game of baseball:

> Even if he makes all the right motions, a man cannot really be said to be playing baseball if he doesn't recognize the umpires' special role in calling players out. (ibid)

[3] Here Walden is appealing to the brute fact/institutional fact distinction drawn by Searle (1969).

In institutional contexts, such as when playing baseball, engagement in the institution requires recognizing the authority of certain individuals … to do what, exactly? Walden is unfortunately not clear here. In baseball, umpires have the authority to determine strikes and balls, whether runners are safe or out, and when a pitcher's herky-jerky motion is a balk and when it is not. Let's say that institutional authorities like umpires have the authority to *determine* a broad range of institutional facts.[4]

Now, Walden wants to say that interpreters are to the institution of interpretation as umpires are to baseball. I take it that what Walden means by this is that interpreters are the ones who determine what count as interpretable.

This analogy would seem to suggest that if we must be sensitive to the ruling of umpires in order to play the game of baseball, then we must be sensitive to the ruling of other interpreters in order to engage in the activity of agency. But does the analogy hold? I think the answer is no. I'm not denying that interpretation is an institution and that there are facts concerning interpretation that are institutional in nature. But unless we are anti-realists about intentions, then we should not believe that the contents of our intentions are institutional facts. Recall that intentions are interpretable *because* they are representational states of the mind, and in being representational states, *their content is public*. It is the *publicity* of states that make them interpretable. But this publicity does not depend on the existence of any interpreters. The publicity of intentions, we can say, is *prior* to the institution that interprets them. The institution that determines the intelligibility of my intentions is one thing, the intentions themselves another. Forming

[4] A strike is an institutional fact. An umpire determines when a strike occurs.

intentions only makes me a *candidate* for an object of the institution – in forming intentions I am now a being who you can interpret – but that doesn't commit me to engage in your institution any more than it commits a chimp or a gorilla.

Of course, if I *do* commit to engaging in your institution of interpretation, then it would seem that I am in turn committed to being sensitive to the authority of interpreters. But what commits me to this institution? According to Walden, it is the fact that interpretability is a condition on agency. But as we've seen, my intentions are interpretable simply in virtue of being public. This is a *brute fact* about my intentions, not an institutional fact, and in no way depends on the authority of interpreters in the way that a called strike is dependent on the authority of an umpire. Satisfying the condition of interpretability would not seem to commit me to anything. In fact, if Wittgenstein is right, it would appear that there is no way I could form a mental state without satisfying the condition of interpretability.

2.4 To morality

Let's set aside the question of whether Walden is licensed to the claim that reflective agents are committed to being sensitive to the authority of interpreters. How does this claim, warranted or not, get us to morality? Answering this question is perhaps the most interesting part of Walden's argument. Walden points out that what we recognize in an individual when we recognize their authority to make claims of intelligibility is their "generic capacities for theoretical and practical reasoning" (p. 95). But this capacity that individuals have to make claims of intelligibility is the very same

capacity that grants people the authority to make second-personal claims of any kind. Walden argues that if I recognize the capacity in you that gives rise to the former authority (to make interpretive claims), then, on pains of incoherence, I must recognize the capacity in you that gives rise to the latter authority (to make second-personal claims of any kind). But recognizing this authority in others, suggests Walden, is what the morality of respect for persons is all about.

2.5 Summary

We can restate Walden's argument as a four-step process:

1. Interpretability is a condition on agency
2. Reflective agency requires sensitivity to the conditions of interpretability.
3. Sensitivity to the conditions of interpretability requires recognition of the second-personal authority of interpreters.
4. Recognition of second personal authority in one domain entails recognition of it in all domains

Step 1 is not problematic when 'interpretability' is understood as 'publicity'. If agency requires exercising an intentional attitude, and attitudes are by their nature public, it follows that all exercises of agency are interpretable. This seems true.

Step 2 is where the argument gets into hot water. If interpretability is understood as publicity, it's not clear what sense we can make of the idea that agency requires sensitivity to the conditions of interpretability, for it is simply impossible to form a state of mind that is not public. But if it is not possible to form a state of mind that it is not public, it's unclear why forming an intention would require being sensitive to the

conditions of publicity, for it will simply be impossible for an agent to form *any* mental attitude that doesn't satisfy the publicity criterion. As such, there is no need to be sensitive to these conditions (since we can't but satisfy them), and it's not clear what such a sensitivity would even amount to.

Step 3 only works if interpretability is understood not as publicity, but as some sort of engagement in an institution of which interpreters are authorities. But, as we've seen, interpretability so understood is not a condition of agency; at least, Walden hasn't given us reason to think so.

I have no problem with step 4. In fact, I'm inclined to think that it is true.

The four steps of this argument can in turn be divided into two parts. The first part of the argument (steps 1-2) concerns the constitutive conditions of agency. The second part of the argument (steps 3-4) concerns the institution of interpretation and its relation to second-personal authority. These two parts are not intertwined. We can accept one part without accepting the second part. This brings us to an interesting parallel with my account. Walden has provided us with a plausible account of why the institution of interpretation might serve as a foundation for morality. I have argued that he has failed to show a constitutive connection between this institution and our agency. In chapter 5 we saw that the version of social constitutivism I was developing in my dissertation ran into a similar challenge.

The real challenge facing the social constitutivist is not necessarily showing that morality is in some way grounded in our sociality – it seems there is a variety of potential options for us – but rather in showing why we are inescapably committed to sociality,

however we ultimately come to understand this. But the sheer plausibility that sociality is an essential feature of the kind of beings that we are, I think makes the thesis an enticing one to continue chasing.

Bibliography

Anderson, Scott. 2017. Coercion. In Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/win2017/entries/coercion/.

Arruda, Caroline T. 2017. Why Care About Being an Agent? *Australasian Journal of Philosophy* 95: 488-504

Ayer, A.J. 1954. Freedom and Necessity. In *Philosophical Essays,* 271-284. London: Palgrave Macmillan.

Bachman, Zachary. 2018. Moral Rationalism and the Normativity of Constitutive Principles. *Philosophia* 46: 1-19.

Barandalla, Ana, and Michael Ridge. 2011. Function and Self-Constitution: How to make something of yourself without being all that you can be. A commentary on Christine Korsgaard's The Constitution of Agency and Self-Constitution. *Analysis Reviews* 7: 364-380.

Bertea, Stefano. 2013. Constitutivism and Normativity: A Qualified Defence. *Philosophical* Explorations 16: 81-95.

Bratman, Michael. 1987. *Intention, Plan, and Practical Reason.* Cambridge, MA: Harvard University Press.

Bratman, Michael. 1999a. Shared Cooperative Activity. In *Faces of Intention*. Cambridge, UK: Cambridge University Press: 93-108.

Bratman, Michael. 1999b. Shared Intention. In *Faces of Intention*, 109-129. Cambridge, UK: Cambridge University Press.

Bratman, Michael. 1999c. Shared Intention and Mutual Obligation. In *Faces of Intention*, 130-141. Cambridge, UK: Cambridge University Press.

Bratman, Michael. 2013. The Fecundity of Planning Agency. In *Oxford Studies in Agency and Responsibility*, Volume 1, ed. David Shoemaker, 47-69. Oxford, UK: Oxford University Press.

Bratman, Michael. 2014. *Shared Agency: A Planning Theory of Acting Together.* Oxford, UK: Oxford University Press.

Bratu and Dittmeyer. 2016. Constitutivism about Practical Principles: Its Claims, Goals and Failure. *Philosophia* 44: 1129-1143.

Broome, John. 1999. Normative Requirements. *Ratio* 12: 398–419.

Broome, John. 2003a. Practical Reasoning. In *Reason and Nature: Essays in the Theory of Rationality,* eds. José Bermùdez and Alan Millar, 85-111. Oxford: Oxford University Press.

Broome, John. 2003b. Reasons. In *Reason and Value: Essays on the Moral Philosophy of Joseph Raz*, eds. R. Jay Wallace, Michael Smith, Samuel Scheffler, and Philip Pettit, 28-56. Oxford: Oxford University Press.

Broome, John. 2005. Does Rationality Give Us Reason? *Philosophical Issues* 15: 321–37.

Broome, John. 2007. Wide or Narrow Scope? *Mind* 116: 359-370.

Broome, John. 2013. *Rationality through Reasoning* (Chichester, UK: Wiley-Blackwell).

Brunero, John. 2007. Are Intentions Reasons? *Pacific Philosophical Quarterly* 88: 424-444.

Brunero, John. 2010. The Scope of Rational Requirements. *Philosophical Quarterly* 60: 28-49.

Burge, Tyler. 2010. *Origins of Objectivity*. Oxford, UK: Oxford University Press.

Burge, Tyler. 2014. Perception: Where Mind Begins. *Philosophy* 89: 385-403.

Chen, Timothy, ed. 2013. *The Aim of Belief.* Oxford, UK: Oxford University Press.

Clark, Philip. 2001. Velleman's Autonomism. *Ethics* 111, 580-593.

Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press.

Enc, Berent. 2006. How We Act: Causes, Reasons, and Intentions. New York, NY: Oxford University Press.

Engel, Pascal. 2013. In Defense of Normativism About the Aim of Belief. In *The Aim of Belief,* ed. Timothy Chan, 32-63. Oxford, UK, Oxford University Press.

Enoch, David. 2006. Agency, Shmagency: Why Normativity Won't Come from What is Constitutive of Action. *Philosophical Review* 115, 169-198.

Enoch, David. 2011. Shmagency Revisited. In *New Waves in Metaethics*, ed. Michael S. Brady, 208-233. New York: Palgrave Macmillan.

Fererro, Luca. 2009. Constitutivism and the Inescapability of Agency. In *Oxford Studies in Metaethics*, Volume 4, ed. Russ Shafer-Landau, 303-333. Oxford, UK: Oxford University Press.

Foot, Philippa. 2001. *Natural Goodness*. Oxford, UK: Oxford University Press.

Gauthier, David. 1986. *Morals by Agreement.* Oxford, UK: Oxford University Press.

Gert, Joshua. 2002. Korsgaard's Private Reasons Argument. *Philosophy and Phenomenological Research* 64: 303-324.

Gilbert, Margaret. 1989. *On Social Facts*. London: Routledge and Kegan Paul.

Gilbert, Margaret. 1996a. Walking Together: A Paradigmatic Social Phenomenon. In *Living Together: Rationality, Sociality, and Obligation*, 177-194. Lanham, MD: Rowman & Littlefield.

Gilbert, Margaret. 1996b. Agreement, Coercion, and Obligation. In *Living Together: Rationality, Sociality, and Obligation*, 281-311. Lanham, MD: Rowman & Littlefield.

Gilbert, Margaret. 2000. *Sociality and Responsibility: New Essays in Plural Subject Theory*. Lanham, MD: Rowman and Littlefield.

Gilbert, Margaret. 2003. The Structure of the Social Atom: Joint Commitment as the Foundation of Human Social Behavior. In *Socializing Metaphysics*, ed. F. Schmitt, 39-64. Lanham, MD: Rowman and Littlefield.

Gilbert, Margaret. 2007. Collective Intentions, Commitment, and Collective Action. In *Rationality and Commitment,* eds. Fabienne Peter and Hans Bernhard Schmid, 258-279. Oxford, UK: Oxford University Press.

Gilbert, Margaret. 2009. Shared Intention and Personal Intention. *Philosophical Studies* 144: 167-187.

Gilbert, Margaret. 2014. Considerations on Joint Commitment: Responses to Various Comments. In *Joint Commitment: How We Make the Social World,* 37-57. Oxford, UK: Oxford University Press.

Greenspan, P.S. 1975. Conditional Oughts and Imperatives. *Journal of Philosophy* 72: 259-276.

Hanisch, Christoph. 2016. Constitutivism and Inescapability: A Diagnosis. *Philosophia* 44: 1145-1164.

Hermann, Barbara. 1996. *The Practice of Moral Judgment*. Cambridge, MA: Harvard University Press.

Hill, Thomas E. 1973. The Hypothetical Imperative. *Philosophical Review* 82: 429-450.

Kant, Immanuel. 1785/2012. *Groundwork of the Metaphysics of Morals*. Trans by Mary Gregor and Jens Timmerman. Cambridge, UK: Cambridge University Press.

Katsafanas, Paul. 2013. *Agency and the Foundations of Ethics*: *Nietzschean Constitutivism*. Oxford: Oxford University Press.

Kerstein, Samuel. 2013. *How to Treat Persons*. Oxford, UK: Oxford University Press.

Kolodny, Niko. 2005. "Why be Rational?" *Mind* 114: 509-563.

Korsgaard, Christine. 1996a, *Sources of Normativity*. Cambridge, UK: Cambridge University Press.

Korsgaard, Christine. 1996b. *Creating the Kingdom of Ends*. Cambridge, UK: Cambridge University Press.

Korsgaard, Christine. 2008. *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford, UK, Oxford University Press.

Korsgaard, Christine. 2009. *Self-Constitution: Agency, Identity, and Integrity* Oxford, UK: Oxford University Press.

Lord, Errol. 2011. Violating Requirements, Exiting from Requirements, and the Scope of Rationality. *Philosophical Quarterly* 61: 392-399.

Lord, Errol. 2014. The Real Symmetry Problem(s) for Wide-Scope Accounts of Rationality. *Philosophical Studies* 170: 443–464.

Manne, Kate. 2013. On Being Social in Metaethics. In *Oxford Studies in Metaethics, Volume 8,* ed. Russ Shafer-Landau, 50-73. Oxford, UK: Oxford University Press.

McCann, Hugh J. 1997. Settled Objectives and Rational Constraints. In *The Philosophy of Action*, ed. Alfred R. Mele, 204-222. Oxford, UK: Oxford University Press.

McCann, Hugh. 1998. *The Works of Agency: On Human Action, Will, and Freedom*. Ithica, NY: Cornell University Press.

McDowell, John. 1994/1996. *Mind and World*. Cambridge, MA: Harvard University Press.

Nagel, Thomas. 1974. “What is it Like to Be a Bat?” *Philosophical Review* 83: 435-450.

Neander, Karen. 1991 “Functions as Selected Effects: The Conceptual Analyst’s Defence” *Philosophy of Science*, 58, 168-84.

Neander, Karen. 2002. Types of Traits: The Importance of Functional Homologues. In *Functions: New Essays in the Philosophy of Psychology and Biology*, eds. Andre Ariew, Robert Cummins, and Mark Perlman, 390-415. Oxford, UK: Oxford University Press.

Nozick, Robert. 1974. *Anarchy, State, and Utopia.* New York: Basic Books.

O’Neill, Onora. 1989. Between Consenting Adults. In *Constructions of Reasons: Explorations of Kant’s Practical Philosophy*, 105-125. Cambridge, UK: Cambridge University Press.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford, UK: Oxford University Press.

Railton, Peter. 1997. On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action. In *Ethics and Practical Reason*, eds. Garret Cullity and Berys Gaut Oxford, UK: Clarendon Press: 53-79.

Rawls, John. 2000. *Lectures on the History of Moral Philosophy*. Ed. Barbara Herman (Cambridge, MA: Harvard University Press).

Rawls, John. 2005. *Political Liberalism: Expanded Edition.* New York, NY: Columbia University Press.

Raz, Joseph. 2005a. The Myth of Instrumental Rationality. *Journal of Ethics & Social Philosophy* 1: 1-28.

Raz, Joseph. 2005b. Instrumental Rationality: A Reprise. *Journal of Ethics & Social Philosophy* 1: 1-20.

Reath, Andrews. 2006. *Agency and Autonomy in Kant’s Moral Philosophy.* Oxford, UK: Oxford University Press.

Reath, Andrews. 2010. Formal Principles and the Form of a Law. In *Kant’s Critique of Practical Reason: A Critical Guide*, eds. A. Reath and J. Timmerman, 31-54. Cambridge, UK: Cambridge University Press.

Roth, Abraham Sesshu. 2003. Practical Intersubjectivity. In *Socializing Metaphysics: The Nature of Social Reality,* ed. F. Schmitt, 65-91. Lanham, MD: Rowman & Littlefield.

Roth, Abraham Sesshu. 2004. Shared Agency and Contralateral Commitments. *Philosophical Review* 113: 359-410.

Roth, Abraham Sesshu. 2014. Prediction, Authority, and Entitlement in Shared Activity. *Nous* 48: 626-652.

Scanlon, T.M. 1998. *What We Owe To Each Other*. Cambridge, MA: Harvard University Press.

Schroeder, Mark. 2004. The Scope of Instrumental Reason. *Philosophical Perspectives* 18: 337-364.

Schroeder, Mark. 2014. The Hypothetical Imperative? In *Explaining the Reasons We Share: Explanation and Expression in Ethics, Volume 1*, 147-172. Oxford: Oxford University Press.

Searle, John. 1990. Collective Intentions and Actions. In *Intentions in Communication,* eds. P. Cohen, J. Morgan, and M.E. Pollack, 401-415. Cambridge, MA: Bradford Books, MIT Press.

Setiya, Kieran. 2007. *Reasons without Rationalism*. Princeton, NJ: Princeton University Press.

Shah, Nishi. 2003. How Truth Governs Belief. *Philosophical Review*: 112: 447-482.

Shah, Nishi and David Velleman. 2005. Doxastic Deliberation. *Philosophical Review* 114: 497-534.

Shpall, Sam. 2013. Wide and Narrow Scope. *Philosophical Studies* 163: 717-736.

Shpall, Sam. 2014. Moral and Rational Commitments. In *Philosophy and Phenomenological Research* 88: 146-172.

Silverstein, Matthew. 2012. Normativity and Inescapability. *Journal of Ethics and Social Philosophy* 6: 1-26.

Silverstein, Matthew. 2015. The Shmagency Question. *Philosophical Studies* 172: 1127–1142.

Silverstein, Matthew. 2016. Teleology and Normativity. In *Oxford Studies in Metaethics*, Volume 11, ed. Russ Shafer Landau, 214-240. Oxford, UK: Oxford University Press.
Smith, Michael. 2013. A Constitutivist Theory of Reasons: Its Promise and Parts. *LEAP*: *Law, Ethics, and Philosophy* 1, 9-30.

Smith, Michael. 2015. The Magic of Constitutivism. *American Philosophical Quarterly* 52, 187-200.

Thompson, Michael. 2008. *Life and Action: Elementary Structures of Practice and Practical Thought*. Cambridge, MA: Harvard University Press.

Tiffany, Evan. 2012. Why Be an Agent? *Philosophical Studies* 90: 223-233.

Valaris, Markos. 2015. The Instrumental Structure of Action. *Philosophical Quarterly* 65: 64-83.

Velleman, J. David. 2000. *The Possibility of Practical Reason.* Ann Arbor, MI: University of Michigan Press.

Velleman, J. David. 2009. *How We Get Along.* Cambridge, UK: Cambridge University Press.

Walden, Kenneth. 2018. Morality, Agency, and Other People. *Ergo* 5: 69-101.

Wallace, R. Jay. 2006. *Normativity and the Will*. Oxford, UK: Oxford University Press.

Way, Jonathon. 2010a. The Normativity of Rationality. *Philosophical Compass* 5: 1057–1068.

Way, Jonathon. 2010b. Defending the Wide Scope Approach to Instrumental Reason. *Philosophical Studies* 147: 213–233.

Way, Jonathon. 2012. Explaining the Instrumental Principle. *Australasian Journal of Philosophy* 90: 487-506.

Wedgewood, Ralph. 2002. The Aim of Belief. *Philosophical Perspectives* 16: 267-97.

Williams, Bernard. 1973/1995. Deciding to Believe. In *Problems of the Self: Philosophical Papers*, *1956-1972*, 136-151. Cambridge, UK: Cambridge University Press.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Trans. G.E.M. Anscombe. New York, NY: Macmillan Publishing Company, Inc.