Ethics and the Possibility of Failure:
Getting it Right about Getting it Wrong

by

David Gordon Dick

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2009

Doctoral Committee:

Professor Peter A. Railton, Chair
Professor Elizabeth S. Anderson
Associate Professor Mika Tapani Lavaque-Manty
Associate Professor Sarah Buss Paulson

Anscombe's Satan can desire evil only by judging it to *be* good, and so he remains, at heart, a lover of the good and the desirable —a rather sappy Satan.

—David Velleman,
"The Guise of the Good"

For

Kent C. Day,
who understood how I might
never finish this,

Gordon H Dick,
who always knew I would,

and

Candace Bolter,
who would have
finished hers long before me.

Acknowledgements

This dissertation marks the endpoint of twenty-five years of official education. In that time I have incurred more debts than I could possibly acknowledge in this space and with my memory. Here, I can only hope to capture most of the people who helped me through the seven years of the Ph.D. program ended by the writing of this dissertation. I will inevitably leave out some who obviously should have been included, and to them I ask their forgiveness, indulgence, and understanding, since they know me well enough to know it was no malice on my part, only the fact that I left this, like everything, to be done in a rush at the last minute.

First I must thank the faculty members who oversaw this project and agreed to be bound to read the whole thing at its completion.

Peter Railton has no shortage of philosophical virtues, but his remarkable talent for seeing the strengths in even the most surprising philosophical views is perhaps his most prominent. Peter's talent at this has made him an unusually good dissertation chair. It allows him to state any view you bring him in its most resourceful form, animating any straw man you might have tried to set up, and reviving hope even when you think your favored view is doomed. This also forces you to consider the arguments on their own merits, because the simple fact that Peter can advocate the view compellingly is no proof that it is correct. I believe that Peter's ability to bring sympathy and strength to any view is what makes him an ideal advisor for a student like me. Peter never dismissed my

claims simply because I could not yet articulate the argument to establish them. Instead, he worked with me to turn enough earth to make apparently infertile land fecund. I have no doubt that what surprising insights I may have generated in what follows could not have found root without this temperament of Peter's. As I have said of him before, the fact that Peter holds a view is fairly strong evidence that the view is correct, but Peter has never sought to train disciples; he trains philosophers.

Through quirks of scheduling, Steve Darwall could not be present for this dissertation's defense and so does not appear on its cover page. This misrepresents the influence he has had on this project and on me. The work on Bishop Butler in Chapter 4 is the direct descendant of a paper written for one of his classes, and is based on some of his own philosophical work. Even more importantly, the entire idea for the narrative arc of the dissertation, moving from justifications to applications of the fallibility constraint, stems from a crucial conversation with him. As would surprise none who know him, no one has had as concentrated an impact on this dissertation as Steve.

Elizabeth Anderson has done a wonderful job of grounding some of my more extravagant flights of philosophical fancy. Her criticisms of and insights on my work have nearly always brought me back to earth where the air is clear enough to see, even when I was lost in the clouds. Liz's emphasis on immediate practical moral problems, in her own work, in her teaching, and when advising me, has been the very model of what a moral philosopher should be concerned about. I am grateful and honored to have had her to influence me.

Sarah Buss came to this project near its end, but her impact on it has been as great as her enthusiasm for it. I cannot thank Sarah enough for always communicating to me

excitement for the questions I was asking, even when I couldn't begin to sketch what my answers to them would be. None of my moments of glum disillusionment with this project were able to survive her exuberant philosophical joy. No doubt some of this excitement came from the fact that Sarah's own work is most closely related to the questions in the first three chapters, and she has been a remarkable resource for me, both by contributing to the literature surrounding these questions in ethics and the philosophy as action, and by orienting me in it. I am incredibly lucky that Sarah came to Michigan when she did.

Finally, Mika Lavaque-Manty has been an ideal cognate member of this committee. Always supportive but never intrusive, Mika's advice about writing, scholarship, and the academy has been extremely valuable. Working with him as an advisor and as a colleague in the CRLT Players was an extremely pleasant way to become acquainted with what is now my ideal of scholarship and engagement. (Although I do hope I can follow his example without also having to take up jogging.)

Though he was not personally involved in much of this dissertation, I must also thank David Velleman for his influence on it and on me. Through his published work, David has provided me with a number of serious philosophical insights; all while showing me that even contemporary philosophy can still be done in elegant turns of phrase. David was also my first advisor at Michigan, and helped me lay the foundation for this dissertation. It was David who first taught me, through his patience and good humor, how important it is to have an advisor who will let you chase an apparently insane idea. I can still see him cocking his head and smiling with surprise to ask me, "That's interesting, why do you think that?" at places where most philosophers would have

simply said “No.” Through his example, David has shown me both what an advisor and a philosopher should be.

Though they were not directly involved in this dissertation and the ideas leading up to it, other faculty deserve thanks for the support they gave me in my time at Michigan. Jessica Wilson was a great first year advisor who became a very good friend to me at Michigan. Ian Proops was a good friend even before he became my advisor in my second year, and remains one of my favorite companions for a drink or a breakfast at Frank’s. Ed Curley was a kind and generous philosopher to teach with, as well as quite a bit of fun to talk with. Henry Dyson did me the favor of tracking me down to talk about Kant and the Stoics, and in the process got me acquainted with him, which was an even bigger favor. Andy Egan turned out to be much more friendly, decent, and simply fun than a philosopher of his stature needs to be. It was been a pleasure getting to know him over beers at Old Town and discussing teaching, philosophy, and everything else. Dan Jacobson has been a wonderful friend and mentor to me while at Michigan. It is just my luck that I am leaving Michigan just as he will be around Angell Hall more frequently, and I might have had the pleasure of his extremely engaging conversation over coffee even more often.

Louis Loeb deserves special mention for all the good he did for me at Michigan. I have never seen a more enthusiastic and engaged teacher than Louis, and I have stolen and poorly imitated much of his signature style in my own teaching. Louis was the first person to seriously introduce me to Hume, the only philosopher I have ever loved reading as much as I loved thinking about. But it was as my placement director that Louis earned his irrevocable place in my heart. Leading more than a dozen Ph.D. students on to the

academic job market in a single year would probably be an impossible task for anyone else, but Louis managed it with a level of dedication, sympathy, and good humor that I will never be able to repay. I can point to at least two specific occasions on which, had I not followed Louis's advice, I would not be heading to employment this summer. Whatever grief I had to suffer from him, and however loudly it might have been delivered, I could never be other than grateful for it because I knew that no one, including my own self, was working harder to get me a job than Louis.

Other faculty members from the University of Utah deserve special thanks, even though I have not seen some of them since I entered graduate school. Bill Whisner was my first philosophy advisor and first ethics teacher, and he did me two incredibly great favors in these two roles. The first was to calm my uptight undergraduate self into believing that getting an A- on my first philosophy paper was not proof that I should give up the study of philosophy. The second was to expose me to the idea that morality could be a matter of happiness and flourishing, not just a matter of resistance and denial. That was quite the revelation to this eighteen-year-old Utah boy. Tom Reed and Yukio Kachi both taught me philosophical and pedagogical lessons that I still try to pass on to this day, and both wrote me letters expressing their confidence in me as I headed off to graduate school. I was touched and honored by their support and hope to one day deliver on the promise they professed to see in me. Bruce Landesman was the first to give me that rarest of things, a job in philosophy, and made me as thrilled as anyone could be to be grading 75 undergraduate papers every few weeks. Nick White and Leslie Francis both pulled me aside and suggested that I should think about graduate school in philosophy, and Leslie was generous enough to let me into her teaching of the proseminar at Utah

while I was still only an undergraduate. Ram Neta and Lex Newman were both wonderful friends to me at Utah, and were both kind enough to support me in applying to various graduate programs.

Even including myself, no one is as responsible for getting me into graduate school as Elijah Millgram. Lije worked with me through eleven drafts of my writing sample, and let me into his seminar on Kant's ethics as an undergraduate. Not only was he crucial in getting me into Michigan, but the basic idea for this dissertation and the first thoughts that led to what is now Chapter 3 were a result of that seminar and my contact with him. Working with Lije for a year was the best preparation I could have had for this Ph.D. program.

I had the enormous good fortune to know Peggy Battin as one of her undergraduate students, research assistants, honors thesis students, and then graduate students. In each one of these roles, Peggy was always my protector and guide, but, more importantly, my friend. Early in my career as her research assistant I saw in her office a bit of microfilm that whose markings made it perfectly clear, that in no uncertain terms, the item absolutely had to be returned to the library two years prior to my birth. Immediately, I knew Peggy and I would get along splendidly, and that there could be a place for people like me in the academy. No one could make the academic life look as attractive as she did, what with her house piled with books, a spare bedroom devoted to each work in progress, and her joyful relationship with her dashing and adventurous English-professor husband, Brooke Hopkins. I have never seen a relationship I wish to emulate as much as theirs, and only members of my own family are as dear to me as they

are. I could not be more grateful for the time I have spent with Peggy and for her influence on me.

Allan Gibbard and Beth Genné deserve special thanks for their friendship and support through my years in graduate school. Through a series of coincidences, I became their default house sitter a mere six months into my graduate career. I spent a total of two of my years in Ann Arbor living in their home, long enough to know where all the kitchen gadgets go and feeling quite at home to wake up there. They provided the location where the vast majority of this dissertation was actually written, as well as the ideal place for its completion to be celebrated. By letting me into their home, Beth and Allan showed me how to live more joyfully even while taking an academic life seriously, and gave me more opportunities to do so. For their warmth, kindness, and continual insistence that it was *I* who was doing *them* a favor, I will always be grateful.

Throughout my time in graduate school I was overseen and helped by wonderful support staff in the philosophy department office. Michele Smyk, Kelly Coveleski, Maureen Lopez, and especially Sue London all did so much to help me, much of which without my knowledge. Molly Mahony was also a great support, but in a more obvious way, since all of her helpful notes and references always did (and will hopefully continue) to show up in my inbox.

As no one in the philosophy department could deny, Linda Shultes is the most capable, most important, and most wonderful person there. Despite having three times as much to do as I did, she always managed to know what was going on with me and to protect me from university bureaucracy, arcane regulations, and my own fallible self. I do not know what I will do without her, and I am not eager to find out.

The vast majority of the financial support for this Ph.D. program was provided by the teaching work I did as a member of my union, GEO, AFT Local 3550. I have worked as a graduate employee both with and without the protection of a union, and so know firsthand how much healthier, safer, and able to work a graduate student is when he is organized in solidarity with his brothers and sisters. If a university is serious about improving graduate education, it will get its graduate students unionized to force it to do what it should have been doing all along.

Beyond the teaching work that supported my time in graduate school, I also had the good fortune to be supported by some non-teaching fellowships, for which I am very grateful. These fellowships were provided by the Philosophy Department, the President's Initiative for Ethics and Public Life, the Rackham Graduate School, the estate of Williston S. Hough, and the Charlotte W. Newcombe Foundation. I was especially grateful and honored to have received the Newcombe Fellowship to support me through the final, and arguably, most difficult year of graduate school.

The Newcombe Foundation provided the financial resources to complete this dissertation, and thereby placed me beyond the ordinary scope of the GEO contract that governed the terms of my employment as a graduate student teacher at the University of Michigan. But it was the standards set by that contract, and the people who fought for it, that provided the psychological resources for it. Following the deaths of my closest family friend and my father, then two major surgeries for my sweetheart, and then the stresses of the academic job market, I fell into a deep depression, unable to do much of anything. In addition to the support of my wonderful friends, it was the access to medical and psychological treatment that has brought me to the point where I could actually write

this dissertation and its acknowledgements. I am especially grateful to Marge Greene for the time she spent listening to my troubles and my jokes. It is simply a mistake to believe that I was afforded this access to care through the abstract generosity of the University of Michigan, since I have watched the access and care expand and improve as the direct result of my union's actions and demands. I am profoundly grateful to these people who fought on my behalf without any knowledge of who I was. As a moral philosopher, I find little more noble than the kind of blind altruism exhibited by so many in GEO.

Through GEO, I found Alyssa Picard, the sharpest mind and the fiercest heart in the labor movement today. I won't embarrass us both by trying to detail all of what she did for me and all I did to her (and her watch) in the more than four years she stood with me as I was writing this dissertation. I will simply express my gratitude that she always did and always will stand behind me, even though we no longer stand together. Thank you for all you did for me, sweetheart.

The other Michigan acronym that was most important to me in my time in graduate school was the CRLT Players. I wish I had found them earlier in my graduate career, since they provided an incredibly welcome outlet and sense of purpose, all while generating some extra income and an opportunity to travel around the country. Never has good work been so much fun. I want to keep working with them whenever it is geographically possible. For all they did for me as a member of their company, I am especially grateful to Emily Wilson-Tobin, Stella Gorlin, Kellyn Webley, Courtney Burkett, and Jeffrey Steiger.

I came to Michigan fearing that its academic climate would be too competitive to foster any real friendships. I could not have been more wrong. From my first dinner

with Soraya Gollop, it was clear that would not be the case. Immediately on arriving at Michigan, I was taken in and treated remarkably well by Steve Petersen, Alex Hughes, Rob Kar, Remy Debes, Carole Lee, and Matt Pugsley. Greg Sax was an especially wonderful companion for conversation over drinks and dessert. Vanessa Carbonell was an excellent ally on the job market, and a good and supportive friend both before and after it. Eleni Manis managed to give me two of the most thoughtful gifts I have ever received, bookending my graduate career. Gabriel Zamosc was a great officemate and good friend, who was always willing to laugh at my jokes, a favor he is hardly stingy with. Marie Jayasekera was the kindest officemate I could have hoped for, especially when I was preventing any of us from working with my shtick. Neil Sinhababu was a surprising addition to the philosophy department, not least for his enthusiasm for Celtic music and sleeping habits, but he turned out to be a dear colleague, friend, and philosophical collaborator. Soraya Gollop threw the finest dinner parties and was the best shopping companion I have ever had, we even found time to talk about philosophy in between and on our outings. Lina Jansson deserves special thanks for all the brunches she arranged for Michael Docherty, Sam Liao, Dustin Tucker and me, and thereby ensured that I would get to know them and her lovely self better. Kevin Coffey deserves similar thanks for organizing a poker night that acquainted me better with him, Steve Campbell, and George Lin.

Outside the philosophy department I was also blessed with an embarrassment of good people. Urmila Venkatesh knows most of them, and all can attest to how fantastic she is. Afia Ofori-Mensa was a great acting colleague and a better friend, and I couldn't be happier that she'll be overseeing my old apartment. As with everything else, she'll do

it more stylishly than I. Gabe Kirchner and Bella Muntz Kirchner are some of the only people who can be forgiven for getting a cat that prevents me from visiting them more often, especially since Gabe's mixes and Bella's one-liners are still coming to me in a steady stream. Victoria Dearman may be the only person who makes video games more fun to watch than to play, which may follow from her general talent at making everything better and more fun. She is the sort of person who will thank you, effusively, for the profound favor you did her in driving her car to Arizona at her expense, and she's so convincing that you'll even believe her for a while. James Mickens, though he is now Vice President of Vice Presidents at Microsoft, should one day have his show about centaurs, so that the world will know the joy that we at Michigan have known for so long which is just to watch him go. Deb Solomon was the greatest neighbor anyone could dream up for the most difficult of times. I wish her the best of luck next year at Harvard, right after she finds it after being misdirected to Shanghai. I wish I had more time to spend with PJ Jacokes, who is that exceedingly rare creature, a genuinely sweet person who is also hilariously funny. Similarly hilarious, but with a somewhat sharper edge, is my friend Joe Davis, whom I'm surprised I don't remember growing up with, since we are clearly brothers. Special thanks to him and to his family, Carly and Tilly, for giving me somewhere to be on New Year's Eve and from here on out. I hope they know that the invitation stands on my side as well. Lesley Braden came along at just the right time to give me a giggle and a lift, and I'm sorry that I won't have more time to spend with her as well. Chris DiAngelo was kind enough to have me out to Go Comedy! twice in one week, and I wish I could have made that a regular occurrence.

In the philosophy department also introduced me to some superb philosophers who are now some of my closest friends. John Ku brought his wonderful mischief into my life at just the right time. Aaron Bronfman tolerated me and joked along with me, all while gently improving my philosophical work. Matty Silverstein, my brother husband, and our mutual benefactor, Bertie, were consistently kind, generous, and playful with me, however far away they might have been. I could not be happier to be attached to them, since ours is a love that cannot parse its name.

Ivan Mayerhofer is possibly the kindest human being I have ever met. It will take a metaphysician of his caliber to solve the puzzle of how he could manage to carry so many of us when he had so much to carry on his own. He is certainly due for a string of good luck, and has evidently found some with Andreea Marinescu, who has been kind enough to share some of her unrestrained joy with me in her wit, laughter, and ability to name colors.

I was first impressed by Alexa Forrester's shoes. Then we discovered, together, and I the hard way, that Michigan ladybugs bite. Later I would offer her a pair of pants. This is a fairly decent way to both summarize and foreshadow my friendship with her, which would later come to include serious philosophy, solid poetry, impressively wonderful parties, and pretty much her entire family. Her own parents have been particularly kind to me, and she brought her own new family to Michigan just in time to give me a home when I needed one most. Her husband, Christian Guenther, turned out to be a good man (and thorough) and has become one of my dearest friends. In addition to recognition of his superior cupcake making skills, he deserves special commendation for putting up with so many philosophers for so long. I should also thank Chris and Alexa

for saving me the hassle of having to choose between them for my favorite member of their family, since they provided the overwhelmingly obvious answer in their son, Cameron Forrest Guenther. Many good things happened to me in graduate school, but on the day I defended my dissertation hearing Cam call me "Doctor Dayes" (he hasn't yet got the "v" sound down) was probably the best.

Joshua Brown and Tim Sundell have become the nearest thing to brothers that I will ever have. (Unless, of course, my parents have been lying to me all these years.) Though Josh persists in his refusal to believe in numbers, despite the fact I *brought him one*, I love him all the same. For all of the patience, intelligence, and good will he offered me in graduate school, I will be forever grateful. Tim has set the bar pretty high for how magnificent a friend can be. Nearly every insightful piece of this dissertation (and most of the rest of my philosophical work) is a direct result of Tim being able to hear what was smart in what I was saying to him. Everything, not just philosophy, is better when you add a Tim to it.

Candace Bolter, the last member of my cohort, managed to provide her love and support to me even after she was no longer around to offer it in person. Candace was the best and kindest philosopher of the seven of us. She came to Michigan to work on political philosophy, but was also a solid philosopher of mind, and had an astonishing philosophical breadth and depth. I remember once telling her a story the punchline of which was how intractable the philosophy of time is. She agreed and commiserated with me, and it was only at her funeral that I learned that she had written an incredibly insightful paper on the philosophy of time, she just hadn't wanted to make me feel bad. Our discipline sorely needs more philosophers like her who are smart enough to crush an

ignorant colleague, but compassionate enough not to. I wish I could see the worlds in which she had her full chance to flourish. They are better than the one we're saddled with now.

My spirits were always kept up by a continual supply of notes, holiday cards, and five-dollar bills from my aunt, Donna Pieper. As a distracted graduate student, I never paid close attention to all the minor holidays, but I'm pretty sure Aunt Donna invented a few just as excuse to send me some of her love through the mail. I'm proud to be her father's namesake, and lament how slow international mail is, since I think it will only delay, not stop, the march of all her wonderful parcels.

Though not strictly related by blood, Ted and Linda Jacobsen might be even better than family, because we got to choose them. Ted and Linda are the anchor of an entire clan of fictive kin that I miss more than anything else since leaving Utah. They are the finest examples of how compassion and sharp wit can coexist (Linda is compassion, Ted sharp wit). I will never be able to detail all they have done for me and how much it has meant, so I hope they will accept my thanks in love instead.

Del and Sadie Ballard are also not related by blood, but just as dear. I am not sure I have ever managed to properly explain to either of them what exactly I was doing out here in Michigan, but they both made it abundantly clear that they were very proud of me for doing it. For their pride in me even as I faltered and for their sincere request for a copy of the first thing I published, I am extremely grateful.

I fell in love with Jim Rounds in our very first conversation (which was, incidentally about GWAR). He turns out to be a remarkably easy person to love, but an even better person to be loved by. Jim has the most magnificent heart I have ever

encountered. He came to visit me in Ann Arbor three times, at serious cost to himself, and has been sure to be with me both as my protector when my stitching was threatening to come apart and as my partner in some of the greatest celebrations that it did not. I think Jim has seen me at my best and my worst, and I hope to keep him around for everything else and in between.

Christopher Howard presents the world with a number of wonderful puzzles. One that confronts me in particular is how he has been one of the greatest supporters of my education while simultaneously being a counterexample to it. After every hard won insight, and every hard fought intellectual battle, I look up to find Chris there, several steps ahead of me, and without a major research university pushing him forward. Through his impeccable eye, clever tongue, and wise mind, Chris has always brought out the best in me, and rather made me feel, by example, that I'm not trying hard enough.

I take pains to mention these features of Chris, because an external observer might miss them in the hurricane of jokes, references, squeaks, and laughter that usually surrounds us. A similar storm of laughter used to surround me when I was with my other friend, Dr. Kent C. Day. Kent first met me when I was three months old and then again when I was 16. I'll let you guess in which of those meetings he indulged my itching desire to talk about Jean-Paul Sartre and theocentric morality. We squandered a perfectly good Christmas party on that one, and a love affair between two nerds was born. Kent was the first academic, non-medical, doctor in my life, and he tried several times to impart his knowledge to me. He wrote the best letters and gave the finest and most thoughtful gifts in either hemisphere. I am physically the caretaker of his watch and his library now that he has given his body to science, but I also hope to carry his style (we

wouldn't be so gauche as to believe in things like spirits) into the future. I am already following his path out of graduate school and into Canada, and I hope to mimic his life in many other ways, except for those he specifically warned me against. It is my second greatest regret that I did not finish this dissertation in time to add a copy to his library before I inherited it.

That library now sits in the basement of my family home where I grew up and where my mother, Yolanda M. Dick, still lives. She has framed every diploma, certificate, and citation I ever earned, including the document proving that I had, at the age of four, successfully completed preschool. I know she thinks I have grown away from her, both geographically and academically here in Michigan, but it was really only when I was away from her that I noticed that all of my best features, the ones that are loved most by me and by others, are just shamelessly stolen from her. She stayed up with me late into the night to copyedit my papers in high school, but she was never an overbearing academic stage mother. It was she who sometimes insisted that I take a break and stay home for a "sick day" now and then when I was too tightly wound, and even my living 1600 miles away did not prevent her from making my lunch and taping a cartoon inside the lunchbox. She is the most fun and the most loyal person I have ever known, and I can only hope that people will always see bits of her in me.

I shared the protection of her ferocious loyalty with my father, Gordon H Dick, who was my best friend, most ardent supporter, and the love of my life. My greatest regret is that I did not finish this dissertation in time for him to see it, even though he forgave me for this before we lost him. I was told by the army of nurses, grief counselors, and friends that what I had to be sure to do was to say to him anything that

needed saying before he couldn't hear it anymore. I embarked on an attempt to do this once a few days before he died, but he waved me off and said, "Just hold my hand." He was right. There was nothing else I needed to hear that he hadn't said, and there was nothing else I needed to say that he hadn't heard. I have learned that the greatest consolation of losing a genuinely magnificent father is that you only regret being unable to tell him new things and see the delight in his eyes and hear the pride in his chuckle; you're only sorry that you can't have more, not that you didn't get enough.

Since I'm paying the fee to send this dissertation to the Library of Congress, and since there is no music to force me off stage, I will end by quoting what I said about my father just before we buried him, and let it live beyond that terrible day.

> If any of you happened to speak with my father in the last thirty years or so, I'm sure you got an earful about me. He made no secret of his pride in me, least of all from me. It was a pride I worked hard to deserve, but I now know it was a pride I could not lose.
>
> And for this, I am more proud of him than he ever could have been of me. He was capable of something that very few people are. As my father, he was capable of loving, completely, a creature he could not understand, completely.
>
> Whatever outlandish outfit, opinion, or piece of jewelry I might show up in, his love for me never flickered or threatened to go out. Even through our fiercest disagreements, he was always on my side, if not of my opinion. More than anyone else, he gave me my philosopher's temperament, which allows me to expect to be perfectly friendly with people I've just spent hours arguing with.
>
> It is no simple trick to love even what you cannot understand, and I have no confidence that I will be able to do as he did, and now I don't have him to set me right.

After all the time he spent with me, studying for exams, talking about papers, strategizing about school, and fighting to give me space and means to pursue my dreams, I know he'd be delighted about how it turned out, and he'd be astounded at my luck.

He'd also wonder why I had to go on so in the acknowledgements, but, as I mentioned, he would forgive even this indulgence of mine. I hope those that read these acknowledgements can too.

Foreword

"There is no normativity if you cannot be wrong."[1] This is how Christine Korsgaard elegantly states the idea that I have taken to calling "the fallibility constraint."[2] I have come to the idea through the study of philosophical ethics, but the consideration also crops up noticeably in the philosophy of mind[3] and the philosophy of language.[4] If Korsgaard's claim is correct, then we should expect the consideration to be relevant, even if it is not explicit, whenever normativity is involved. It is not my project here to give an exhaustive list of all the instances where the fallibility constraint comes into play, be it implicitly or explicitly, in philosophical thinking about the normative. However, as Korsgaard states it, the fallibility constraint's intuitive pull and universal scope makes us expect to find it operating, even if only covertly, in all our discussions of what is normative and of normativity itself.

Reversing Korsgaard's formulation of the claim to the positive thesis that "There is normativity only if you can be wrong," better highlights its ambiguity. Insisting only that something "can be" wrong instead of "is" wrong is a claim about possibility, and to understand it we have to settle the question of what sort of possibility is involved here. As it turns out, disambiguating this claim will take up a significant portion of the

[1] Korsgaard, 1996, page 161. Hereafter referred to parenthetically in the text as TSN with page number(s).
[2] Douglas Lavin calls it the "error constraint." See Lavin, 2004.
[3] See Fodor, 1987.
[4] See Kripke, 1982, and Rosen, Manuscript.

following dissertation. I will argue that some interpretations of it cannot possibly be correct, if they are to be compelling and applicable to moral philosophy. What implications this will have for other philosophical thinking about other normative enterprises (such as epistemology and semantics) I will only offer a few concluding remarks and then leave to later consideration.

As Korsgaard's statement of it reveals, the fallibility constraint is meant to be a perfectly general feature of normativity itself, and so its truth and applicability will not be restricted only to ethics, but since I have come to this problem through moral philosophy I would like to motivate it here with a brief sketch that comes from moral philosophy. This comes from Kant's philosophy of action.

What gives an action moral worth, according to the storybook version of Kant I was first exposed to as an undergraduate, is that it is autonomous.[5] In addition to bestowing moral worth, autonomy is also what makes an action an action that is fully authored by its agent instead of an instance of that agent's mere behavior. The opposite of autonomy is heteronomy, when my will is given a law from some force alien to itself rather than being given a law from its very self, as the words' etymologies suggest. Whether the outcome is good or ill, when I behave heteronomously it is an act[6] with no moral worth and (or perhaps because) it was not really an action I fully authored. What caused the act was the force outside my will that took over, not my will itself. When it is my will that gives itself a law and is therefore autonomous, then my action will both have

[5] I here call this version "storybook" so as to flag my wish to bracket all questions of Kant interpretation here. Whether or not the historical Kant's account could meet the demands of the fallibility constraint, it is how the fallibility constraint is functioning here that matters for present purposes.

[6] Please note that I am here and hereafter following W.D. Ross and using "act" and "action" as technical terms. The complete description of action includes both a doing and its motives or aims. So, when I buy you ice cream to help you gain weight after your illness, I perform the same act as your friend who buys you ice cream to soothe your tonsillectomy wounds, but a different action. See Ross, 2002.

moral worth and be a genuine instance of me acting, instead of being overtaken by usurping alien forces.

But what about evil? Not just when good intentions go awry due to ignorance, circumstance, or simple bad luck, but the genuine evil of an agent who acts on his malicious motives and understands the badness of those motives. It is the evil of an agent who sets out to make you suffer, precisely because that suffering is bad and he wishes to do you harm. Surely this sort of rationally considered misbehavior occurs, but it seems that on this version of Kant, it cannot. When my acts lack moral worth, i.e. when they are heteronomous, it is because they are not really my doing, they are the result of other forces overtaking me. When my acts have moral worth, i.e. when they are autonomous, they are both mine and morally worthy, biconditionally. My actions have moral worth if and only if they are mine. Putting the point this way might conceal what is troubling about this biconditional, but reversing it should help. If and only if they are mine, do my actions have moral worth. This means that whenever *I* act, I act well. There simply is no way for me to direct my will in a morally unworthy way. The only way for this to happen is for my will to not be mine.

At this point, some readers, my undergraduate self included, will think that something must have gone wrong.[7] Furthermore, it will seem that something perfectly general and quite obvious has gone wrong. Precisely what has gone wrong is still obscure, but some worries suggest themselves already. How could there be morally worthy actions if there are no morally unworthy ones? Does this mean that Kant rejects

[7] There is good reason to think that such readers include Sidgwick, Schelling, and Kierkegaard in addition to David Dick in college. See Sidgwick, 1888, and Kosch, 2006.

the possibility that people could knowingly do wrong? Isn't it just obvious that there is no normativity if you cannot be wrong?

For some time, even after I finished being an undergraduate, I was content to take the mere observation of this problem as a decisive consideration against Kantian ethics. As we shall see in Chapter 2, it is a popular enough move to reject a moral philosopher's views on the grounds that they cannot meet the demands of the fallibility constraint.[8] And it is still a consideration that I think is decisive against at least some Kant*ian* accounts of morality, as we shall see in Chapter 3. But it is not a consideration that is beyond the resources of all philosophers in the broader autonomist tradition of which Kant is a part, as I aim to show in Chapter 4.

This dissertation is also concerned with the questions that are presupposed by the application of the fallibility constraint to any particular moral theory. In Chapters 1 and 2 I seek answers to the questions of *why* and *how* the fallibility constraint does and should govern our thinking about ethics. Chapter 1 examines and rejects a seemingly obvious way of vindicating the demands of the fallibility constraint by appealing to the nature of rules. As Wittgenstein observed, it is not a rule unless you can break it, and moral rules should be no different, so this would vindicate the fallibility constraint, were it correct. Chapter 2 also involves rejecting another promising basis on which to justify the fallibility constraint, the Principle of Alternate Possibilities, but offers in its stead a more stable grounding, rooted in the reactive attitudes that are an ineliminable part of our moral life and practice.

[8] I have in mind specifically Korsgaard's arguments against Hume in "The Normativity of Instrumental Reason." See Korsgaard, 1997. Hereafter referred to parenthetically in the text as NIR with page number(s).

Though it was this collection of the questions from Kant's own moral philosophy that first motivated this dissertation, in it I will remain silent on whether or not those considerations constitute fatal objections to Kant's theory, or even if they are reasons to modify it beyond recognition. This is a matter I will leave to stronger scholars of Kant and the German language than my current self. Even though this dissertation will not settle the particular question that sparked it, I hope it will make progress toward settling some other questions in moral philosophy, particularly by increasing our understanding of the strength and nature of this tool, the fallibility constraint, that we might later use to settle questions in moral theory more decisively.

Table of Contents

Abstract

Ethics and the Possibility of Failure:
Getting it Right about Getting it Wrong

by

David Gordon Dick

Chair: Peter A. Railton

Entire moral philosophies have been rejected for ruling out the possibility of failure. This "fallibility constraint" (also sometimes called the "error constraint") cannot be justified by appealing either to Wittgensteinian considerations about rules or to the moral importance of alternate possibilities. I propose instead that support for such a constraint in ethics can be found in the Strawsonian reactive attitudes. I then use the constraint to reveal hidden weaknesses in contemporary contstitutivist strategies to ground moral normativity such as Christine Korsgaard's, and also to reveal hidden strengths in historical accounts of morality such as Bishop Butler's. We will have reason to reject any moral theory that makes constitutivism's mistake, but only because we have reason not to reject the fallibility constraint itself. The way this ethical fallibility can be justified suggests a general principle that could be used to justify fallibility constraints in other normative domains such as practical reason, epistemology, the philosophy of language, and the philosophy of mind.

Chapter 1

The Normativity of Unbreakable Rules

Whatever normativity is, it seems that it must be bound up with possibility of failure. To make the normative observation that things are not as they ought to be is to notice a failure in the world. To observe that things are just as they should be seemingly adds nothing except in contrast to how things might, but should not, have been.

A world in which nothing ever went wrong would be remarkably different from our own, in a variety of ways. Hardly the most remarkable but, for present purposes, the most relevant feature of such a world is that it might not even have a notion of normativity in play. Almost certainly such a world would not also include humans, but if it did contain some such infallible sentient beings they might never have bothered to develop anything resembling our notion of normativity.

The reason why might depend on one's view about how this imagined world came to have no failure in it. If it is a world where nothing has ever failed as a matter of the most fantastic coincidence, then we will probably wish to claim that in this case the world is a hospitable habitat for normativity, it is just that the sentient beings populating it never had need to develop a notion of the normative, and so have not observed a feature of their world that is present, just obscured by all the success. Just as we might never discover the colors of the autumn leaves if our climate is so temperate as to prevent the green from ever shrinking from the leaves in the cold, these sentient beings might never

discover that there is a way that things *should* be, since they have never been confronted by things as they should not be. It might be a pleasant pastime for these beings to observe that all is as it should be, but there is nothing in the world forcing them to notice it. Presumably, their luck may also keep them ignorant of other things like disappointment and frustration.

Alternately, if this world has achieved its flawless perfection as a result of some kind of *necessity* (which kind need not concern us at the moment) we might be tempted to claim that such a world can have no such thing as normativity in it at all. A world at risk for containing failures but that never suffers them might make the notion of normativity *invisible*, but a world that contains no failures because it was never at risk to contain them might make normativity itself *impossible*. Kant himself apparently thought this was true, as we find in the *Lectures on Ethics* the remark, "[O]nly God's will is automatically good and perfect, and we cannot say of Him, as we do of men, that He ought so to act."[1]

This is just the sort of thing we should expect if we read pronouncements like Christine Korsgaard's "There is no normativity if you cannot be wrong," (TSN 161) as deep metaphysical commitments. If the possibility of failure is closed off, so too, it seems, is the possibility of normativity. Beings in a world where it is simply not possible to fail might have no normativity to perceive, whether or not they developed a concept of it.[2] Conversely, if the possibility of failure persists and is simply never realized, beings

[1] Kant, 1997. Academy page 29:605, Cambridge Edition page 230. This remark was recorded by C.C. Mrongovius.

[2] It is also an interesting question if a world that necessarily contains no failures could contain any sentient beings at all. It is a popular enough move to claim that angels are creatures that are free but never sin (see Plantinga, 2007, p. 325 and Herman, 1993, pp. 59-60), so let me, for the moment, suggest without argument that such a necessarily flawless world could contain angels who might have the concept of the normative, even if their world has nothing normative to which to apply it.

witnessing such good fortune might be able to observe that all is as it should be, if they managed to develop and apply such a concept.

The idea that there simply can be no normativity when there is no possibility of failure is an influential one in contemporary philosophy.[3] It is not within my interests or current capacities to make an exhaustive historical study of its genealogy, but there are some obvious sources for the idea and its influence on the philosophy of the last thirty years or so.

I. Kripkenstein, Rules, and Normativity

To many readers in my philosophical tradition, Saul Kripke's *Wittgenstein on Rules and Private Language*[4] might spring to mind as an obvious source of this idea and its pervasive influence. There, Kripke practically provides a definition of the normative as something involving failure. In arguing against a dispositional solution to the puzzle of how I might be sure that I mean 'plus' by the symbol '+' instead of some deviant function like 'quus,'[5] Kripke says

> The dispositionalist gives a *descriptive* account of this relation: if '+' meant addition, then I will answer '125' [when faced with '68 + 57']. But this is not the proper account of the relation, which is *normative*, not descriptive. The point is *not* that, if I meant addition by '+', I *will* answer '125', but that, if I intend to accord with my past meaning of '+', I *should* answer '125'. Computational error, finiteness of my capacity, and other disturbing factors may lead me not to be *disposed* to respond as I *should*, but if so, I have not acted in accordance with my intentions. The relation of meaning and intention to future action is *normative*, not *descriptive*. (WRPL 37)

---

[3] See Douglas Lavin's list of quotations of authors asserting things to this effect. Lavin, 2004, pp. 424-425.
[4] Kripke, 1982. Hereafter referred to parenthetically in the text as WRPL with page number(s).
[5] I suspect that anyone bothering to read a philosophy dissertation is familiar with this example, but if not, here is a brief gloss. "Plus" is our ordinary addition function, just as we were taught in school. "Quus" is identical to plus up to any value under 57, when the "quaddition" function will give the answer "5" instead of the usual additive sum. Part, but not all, of the worry here is how I could ever know, prior to hitting values of 57 and above, whether I was adding or quadding. For more on this see chapter 2 of Kripke, 1982.

In short, the normative force of the meaning of '+' consists at least partially in the fact that there is an answer I *should* give to the problem '68 + 57' even when I fail to do so. Normativity persists through failures in a way that descriptive accuracy does not. According to the normative mathematical rule of addition, I should get 125 when adding 68 and 57, even when it is descriptively false that I did. Even though I might actually give the answer I should, the whole concept of normativity seems to ineliminably involve the idea that there is an answer I *should* give even if I fail to do so. Normativity's distinctive ability to survive failures seems to be part of its very nature and the reason why it might only be worth considering in contexts when the world is at risk for such failures to occur. Thinking about normativity will involve thinking about failure.

Kripke's emphasis on Wittgenstein for these ideas combined with his influence on contemporary philosophy likely explains much of why this idea has gained so much currency, and why Wittgenstein seems like an obvious source of it.[6] Indeed, the passages of Wittgenstein that Kripke quotes not only emphasize the relationship between normativity and failure, but they suggest a relationship that looks to just *be* a fallibility constraint, i.e. the idea that there is no normativity if you cannot be wrong. As Kripke claims

> Nothing is more contrary to our ordinary view — or Wittgenstein's — than is the supposition that "whatever is going to seem right to me is right." (§258). On the contrary, "that only means that here we can't talk about right" (*ibid.*). (WRPL 23-24)

In another passage that Kripke quotes as stating "the Wittgensteinian Paradox" Wittgenstein himself provides what might be the *locus classicus* for the fallibility

[6] I suspect that this currency and influence is also due to Wittgenstein's influence on Anscombe, and her influence on us, particularly in works like *Intention* (Anscombe, 2000). But at the moment, I must leave this suspicion substantiated by nothing more than my hunch that it is true.

constraint. On his way to provoking a worry that there might just be no such thing as following a rule, Wittgenstein remarked:

> This was our paradox: no course of action could be determined by a rule, because every course of action can be made to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here.[7] (PI 201)

Wittgenstein's point here was not merely to separate the legitimate rules from the illegitimate ones. The surrounding remarks indicate that every rule involves an interpretation and a practice to obey it, and perhaps since a rule is subject to limitless interpretations and practices it might be true of any rule that "everything can be made out to accord with the rule."

Remarks like these have led to worries that it might be impossible to determine, for any act, which particular rule the agent performing the act might be following, and so there might just be no such thing as following a rule.[8] Even though, on this interpretation, Wittgenstein challenges the very possibility of rule following, Remark 201 indicates a feature of rules that could separate the illegitimate from the legitimate ones, if there were such things.

In fact, Wittgenstein's statement in Remark 201is an excellent candidate for the fallibility constraint as it is stated and used in contemporary moral philosophy. It demands that a legitimate rule must be capable of giving guidance. To give that guidance there must be things that count as breaking the rule, because if everything counts as following the rule, then all courses of action are recommended equally and so would

[7] Wittgenstein, 1998, Remark 201. Hereafter referred to parenthetically in the text as PI with reference to remark numbers.

[8] The worry runs deeper than just an epistemic difficulty, at least as Kripke understands it. Kripke himself says as much: "Recall that the skeptical problem was not merely epistemic. The sceptic argues that there is no fact as to what I meant, whether plus or quus." (WRPL 38)

appear to make the rule impossible to follow. The idea is that nothing can be determined to be an instance of following the rule if there are no potential circumstances of rule breaking to compare against them. In short, Remark 201 claims that there is no normativity if you cannot be wrong.

Furthermore, Wittgenstein's formulation of this maxim seems to be powerful and obviously correct, almost analytically so. Wittgenstein's observation seems to be that once we understand what a rule is, then there will have to be something that counts as being wrong with respect to the rule if there is to be something that counts as being right with respect to the rule. Understood this way, it is a consideration that could well be used to support the work in contemporary moral philosophy that tends to rest upon it. There, philosophers argue against opposing moral views by working to show that the moral theory under examination cannot be the correct one because it proposes rules that are unbreakable and so not normative.[9]

To see how Wittgenstein's considerations might be used in this way to disqualify a moral theory as illegitimate, we can use it to generate a novel objection to Divine Command Theory. The point of this exercise is to give an example of how a Wittgensteinian version of the fallibility constraint could be used to argue against a moral theory. Pitching it against Divine Command Theory seems especially appropriate, since, as perhaps the least plausible moral theory, we will not have reason to doubt it if our argument concludes that Divine Command Theory is mistaken. Of course, I mean to hang no serious argumentative weight on this argument, and so even if a reader should be convinced that Divine Command Theory is correct, she can still appreciate the way the

[9] I have in mind here pieces like Christine Korsgaard's argument against Humeans in "The Normativity of Instrumental Reason," (Korsgaard, 1997). See Chapter 2 for my discussion of it.

fallibility constraint is being applied to the theory, even if she remains unconvinced that it disproves the theory.[10]

Divine Command Theory holds that something is right (or good) just in case it is God's will. Already, fallibility worries are lurking in the account since, obviously, God cannot do anything wrong, since, as soon as He wills it, it becomes right. This much is not yet a violation of the fallibility constraint since it is still possible for someone (i.e. anyone who isn't God) to do wrong, should they do anything contrary to God's will. Anyway, it is hardly comes as a surprise if God, the perfectly good being, is unable to do anything wrong.[11]

But we can get Divine Command Theory to run up against fallibility worries if we recapitulate one of the standard objections to it.

Ordinarily, one can illustrate a problem with Divine Command Theory by pointing out that if *anything* God wills thereby becomes right, then God could will a particularly morally abhorrent thing and thereby make it right. By showing that Divine Command Theory would entail that God could make, say, malice toward innocents morally obligatory simply by willing that it be so, we seem to have found our reductio of the theory. It seems to contradict the very idea of 'right' to suggest that something so awful could be right, even if God wanted it to be, and so Divine Command Theory must be mistaken if it has such an entailment.[12]

---

[10] As it happens, I am convinced that Divine Command Theory is wrongheaded, but, as I will show at the end of this chapter, I am not convinced by the argument I am about to give.

[11] It is a good but tangential worry that this inability might limit God's omnipotence, so I won't take it up here.

[12] Interestingly enough, there seems to be no thing so awful that everyone agrees that God could not make it right. Compelling examples tend to differ from person to person.

The way Divine Command Theory might violate the Wittgensteinian fallibility constraint is structurally analogous to the difficulty just outlined, but instead of plugging in something morally abhorrent as God's will, all we must do is plug in something that is infallible, in the sense that concerns Wittgenstein and Kripke. To take an example that both would agree generates the troublesome sort of infallibility,[13] suppose that God were to will it so that right becomes "whatever seems right" to the agent performing the action. There are numerous difficulties with the theory proposed in this suggestion, but the Wittgensteinian understanding of the fallibility constraint provides a particularly elegant objection, and one that appears to be logically prior to many of the other available complaints. The objection is that the principle "whatever seems right to an agent is right" cannot be the fundamental moral rule, because it cannot be a normative rule at all. If agents can only ever do what seems right to them, then they cannot possibly violate this rule. Anything they do will be what seemed right to them, and so nothing they can do can count as being in violation of this rule. If this is the case, this rule cannot even be normative in the first place and so *a fortiori* could not be the fundamental moral rule, even if God willed it to be. Since Divine Command Theory entails the possibility that rules that cannot be normative are normative, we have found another reductio to bring to bear against Divine Command Theory.

In both versions of this objection Divine Command Theory is rejected on the grounds that it would admit a contradiction. In the standard version, the worry is that God could make what is manifestly wrong right. In this new version based on the

[13] See Wittgenstein, 1998, Remark 258 and Kripke, 1982, pp. 23-24. As I will show later in this chapter, this is not, in fact, an infallible rule, but I have chosen it because both Wittgenstein and Kripke cite it as an example of one, and so using it makes this example of the use of a Kripkensteinian interpretation of the fallibility constraint as orthodox as possible.

Wittgensteinian fallibility constraint, the worry is that God could make normative what manifestly could not be normative. In either case, Divine Command Theory threatens to eviscerate a central moral concept, and so, as the argument goes, cannot be the proper analysis of morality.

In this example, the Wittgensteinian version of the fallibility constraint is doing just what we would require of a fallibility constraint in moral philosophy. It is providing us with novel reason to abandon a moral theory as mistaken, and it is doing so for precisely the sorts of reasons that Kripke and Wittgenstein observed. Because Divine Command Theory could propose a moral rule according to which nothing could fail to be right, "that only means that here we can't talk about 'right'." (PI 258)

We have now at least established that the Wittgensteinian understanding of the fallibility constraint can be employed in just the way a fallibility constraint in moral philosophy should be. If Wittgenstein's observations about the nature of rules are correct, we will have found a fallibility constraint applicable to moral philosophy, along with a vindication of it.

As I will argue in the next section, however, it is not yet time to stop searching for the proper analysis and proof of the fallibility constraint because Wittgenstein's claims about the nature of rules are mistaken. By examining a number of potential responses to a particular unbreakable rule, I hope it will be revealed that normativity (specifically the ability to follow a rule and to be evaluated in terms of it) is not confined to only breakable rules. Once this has been established, I will return to the private language argument and the potential threat it poses to this counterexample, and then conclude the chapter with a real practical problem (i.e. deliberative paralysis) suggested by, but not

endemic to, this counterexample, which is one that I take to be decisive proof against the Wittgensteinian analysis of the fallibility constraint.

II. An Avant-Garde Counterexample

If the Wittgensteinian claim about rules were right, then it should be impossible to properly derive guidance from an unbreakable rule. Furthermore, if only breakable rules could be normative, then it should be impossible to evaluate performance in light of unbreakable rules. I hope that the following thought experiment will show that neither of these claims is true.

To illustrate this, imagine that an art instructor decided to conclude his course by giving an assignment in avant-garde art. This final assignment is to turn in anything, including nothing at all.

We may be tempted to say any number of disparaging things about this assignment, and about the instructor who gives it. But the question for us is whether the instructor has given an assignment that provides no guidance and so is not normative.

Before we try to determine whether or not this assignment can provide normative guidance, it will be helpful to have some idea of what normative guidance is. In a recent paper entitled "Normative Guidance," Peter Railton has provided an insightful and useful, though deliberately not exhaustive, sketch of what it is to be normatively guided.[14] There, before he further refines his definition to include less conscious normative guidance, Railton offers this description:

> Conduct C is guided by norm N only if C is the manifestation of a disposition to act in a way conducive to compliance with N, such that the

[14] Railton, 2006. Hereafter referred to parenthetically in the text as NG with page number(s).

> fact that C conduces to compliance with N plays an appropriate role in the explanation of the agent's C-ing. (NG 8)

While Railton is discussing norms and not rules, we can show that a rule can function as a norm and offer the guidance Railton describes. To determine whether the avant-garde art assignment can be normative, we should examine some of the ways students might conduct themselves and see whether this conduct can genuinely play the "appropriate role in the explanation" of the student's "disposition to act in a way conducive to compliance" with this assignment. To begin, consider two students, Ernestine and Albert.

On hearing the assignment, Ernestine, who is majoring in art and in particular studying sculpture, thinks to herself, "Great! I'll finally have an excuse to make that sculpture I've been thinking about all year." Ernestine then sets about working on her project and on the day the assignment is due, brings in her precisely scaled replica of Duchamp's *Fountain*, rendered in Campbell's Tomato Soup cans.

Ernestine's classmate, Albert, is also excited when he hears what the assignment will permit him to explore. Albert thinks to himself, "Excellent, I'll finally have a chance to express all this emptiness and nothingness I've been feeling. In fact, what better way to express that than by turning in *nothing at all*? In fact, I won't even show up! My absence will be deafening."

Whether or not Albert is correct about the impact his submission will have, his implicit assumption that it will count as a way of satisfying the demands of the assignment is exactly right. This is no surprise, since the assignment itself has provided him with all the guidance he needed to fulfill it.

We can reconstruct Albert's thought process (implicit or explicit) like this: "I have to complete my assignments, because I need to finish college in order to get the job

I want. This assignment on avant-garde art is one I need to complete, and I see it offers many equally good ways to complete it. To break this tie, I'll pick a way of completing it that satisfies one of my other needs or desires. So I'll turn in nothing as a way to express this feeling I want to express."

Granting that Albert has some thought process like this and that it is what leads him to submit nothing for his art assignment, it seems we have a clear case of normative guidance, as Railton describes it. Albert's submitting nothing does reflect both his disposition and intention to act in a way that is conducive to complying with the requirements of the avant-garde art assignment. Furthermore, it is a fact that Albert's submission of nothing is conducive to complying with the assignment, and this fact plays just the right role in the explanation of Albert's submitting nothing. Albert wants to fulfill his assignments and sees that, in this case, submitting nothing is a way of fulfilling the assignment. It is his understanding of this fact (among other things) that leads him to submit nothing.

Not only has Albert found a way to successfully complete his assignment with enviably little work, he has performed an even more impressive feat. He has just done what Wittgenstein thought to be impossible, since he has managed to follow a rule that he cannot violate. Even though it turns out that whatever Albert had submitted would have completed the assignment, this does not mean that Albert could not derive genuine normative guidance from the rule governing the assignment. He understood —correctly— that submitting nothing counted as a way of completing his assignment, and this understanding helped guide him to do precisely that in response to the

assignment. The assignment gave a rule he could follow, even though it was a rule he could not break.

For those worried that there is something slippery or anomalous about Albert's case that makes it illegitimate, it is worth a moment to consider Ernestine's case alongside it. Ernestine's thought process could be reconstructed in just the same way as Albert's, except that her additional preferences led her to break the tie among the various permitted options by making a sculpture she has wanted to make. This makes Ernestine's case relevantly similar, but with one very important counterfactual difference.

Had the assignment instead been to turn in anything at all *with the exception of* turning in nothing, there would be no question that Ernestine's thinking and sculpture would have counted as appropriate responses to an assignment containing a perfectly legitimate (however pedagogically degenerate) rule. It is especially important to notice here that it would still be a case of legitimate normative guidance even though Ernestine's decision would be normatively underdetermined, in that the rule would give her multiple, equally permissible options from which to pick.

Of course, Albert's thinking and submission would not have satisfied this alternate assignment, and, for those who think only breakable rules can be normative, it is this fact that somehow shores up the claim that this alternate assignment had a genuinely normative rule while the original assignment did not.

How this fact might do this, though, is quite mysterious. By permitting artistic submissions of nothing and thereby making the rule in the assignment unbreakable, the instructor seems only to have given students like Ernestine one more permissible option in an already wide selection. This extra option will also make thoughts and submissions

like Albert's count as appropriate responses to the assignment, but in doing so it has not made Ernestine's guidance merely fictional. She would be quite right to think that her sculpture would satisfy either assignment she might have been given, because either assignment permits such a sculpture to count as an appropriate submission.

The fact that Ernestine's sculpture would satisfy the demands of either assignment can be further illustrated if we consider how the instructor who gave this unusual assignment might respond to the submission in each case. This will also give us a chance to consider another aspect of normativity, that of normative evaluation.

Beginning with the uncontroversial case, suppose that our instructor gave the assignment that permits submitting anything with the exception of submitting nothing at all. When he receives Ernestine's sculpture, the rule in the assignment will easily allow him to judge how well it measures up. Considering her submission against that rule, the instructor will be able to notice that, as a sculpture made out of soup cans, it certainly counts as something and so deserves credit. The instructor might even accurately think to himself, as he recorded the credit she had earned, "Good thing Ernestine turned in something rather than nothing, or else she would deserve no credit."

This thought would not be accurate had the instructor given the other assignment that would credit submissions of nothing at all. Even though this further thought would not be appropriate in the context of this other assignment, this does not imply that any evaluation would be similarly inappropriate. When considering Ernestine's sculpture as a submission to satisfy the assignment that accepts anything, including nothing at all, the instructor can still correctly discern that the sculpture counts as one of the many ways of completing the assignment. To deserve credit for this assignment, students need only

turn in what could be found on the list of anything, including nothing at all, and Ernestine's sculpture is certainly included there. Even though there is nothing Ernestine could have done to fail to deserve credit for this assignment, this will not mean that what she ultimately does will not deserve credit. Surely there is something wrong with this assignment, but the problem is not with its capacity to offer guidance or permit evaluation.

Wittgenstein's explicit claim was that "if everything can be made out to accord with the rule, then it can also be made out to conflict with it." (PI 201) But the avant-garde art assignment shows us that this is not true. It is a case where, since everything can be made out to accord with its rule, then *nothing* can be made out to conflict with it. Unbreakable rules like these turn out to provide an opportunity for universal compliance, rather than being standards according to which there can be "neither accord nor conflict."

At this point, I take myself to shown that a rule can be followed even if there is no possible way for it to be broken. But in the example of the avant-garde art example, I have not managed to dispel every relevant possibility of failure, and these should be examined before I conclude my presentation of this case. To do this, I will present four more students to illustrate further potential sources of fallibility, but none of which substantiate Wittgenstein's claim that, relative to an unbreakable rule, "there would be neither accord nor conflict here."

First, consider Jim, who, like Albert, also is absent and turns in nothing on the day the assignment is due, but not as a nihilistic artistic statement, but rather as a result of Jim's sheer slothfulness and neglect. After enrolling in the class, Jim began spending many late nights out and frequently slept through his alarm or simply forgot to set it to

wake him in time to make his art class. As a result, Jim missed the day the avant-garde art assignment was given and also slept through the class on the day it was due. Again like Albert, Jim will receive full credit for his sleeping through the class on the assignment's due date, but here it seems that something has gone wrong. Albert skipped out on class as an artistic statement; Jim is making no conscious statement about the assignment through his absence, he cannot, he is not even aware that the assignment has been given. This seems like a kind of relevant failure that Albert could have committed, even though his submission would have been the same. So why isn't this the sort of failure that is necessary for normative guidance and what vindicates the Wittgensteinian claim?

To answer this question, let me draw a distinction between *following* a rule and *satisfying* it. To follow a rule, one must do as Albert has done, and organize his thinking around the rule and its demands, perhaps with an aim to doing as it demands. Often, agents follow a rule with an aim toward satisfying it, but there are some rules one can satisfy accidentally, without even attempting to follow them.[15] In keeping with scholastic cases, consider a multiple-choice exam. There, there is a rule instructing the students to circle the correct answer to the given question. One student might actually follow this rule, by dutifully reading the question, carefully considering it and then selecting the correct answer to circle. Contrast this with a student who has a series of seizures through the exam that prevent him from reading the question, but nevertheless result in his circling the correct answers. Here, both students have *satisfied* the rule of this multiple-choice exam, but only the first student has done so by *following* the rule. It is certainly a

[15] As we shall see shortly, I'm willing to admit that organizing one's thinking around a rule in order *break* it is also a kind of following, but I think it is merely a semantic quibble about whether this should be included in the definition of following itself.

limitation of multiple choice exams that they will count as correct answers circled as a result of seizures or merely lucky guesses, but no one should deny that such circlings nevertheless satisfy the demands of the exam. They just don't satisfy the demands in the way we might have hoped.[16]

When Jim sleeps through class, ignorant that the avant-garde art assignment has even been given, the structure of the assignment will ensure that he satisfies its demands even though he has not followed its rules. This may well be a pedagogical limitation of this assignment, but it is not enough to undermine the fact that it is an assignment that contains a genuine rule.

Further, the possibility that Albert might have failed to follow the rule as Jim has does nothing to alter the fact that they both count as having satisfied the rule and that Albert has managed to follow it. Wittgenstein's claim was that a rule according to which everything counts as satisfying can be neither satisfied nor followed because there can "be neither accord nor conflict" in such a case. Indeed, a student might fail to follow the rule, as Jim has done, but what the avant-garde art case shows is that this does not entail that no one else could either follow or satisfy the rule, as Albert has done.

But how is it that Jim counts as satisfying the rule through his ignorance of it and failure to respond to it in any way? Won't this prove too much and have it so that every student at the college satisfied the rule, whether enrolled in the class or not? In fact, if

[16] If one is inclined to object that a student suffering from seizures nevertheless counts as *satisfying* the demands of the rule that instructs him to circle the correct answer to the given question, allow me to offer the following. By "satisfy," all I mean is something like "behave in a manner consistent with the requirements of the rule" even if this behavior is purely accidental, and occurs without knowledge of the rule. For those uncomfortable with the seizure case, consider a student who arrives at school dressed in a jacket and tie for a formal event held immediately after classes that day, but who is surprised to learn that he fits right in on the school's "Formal Dress Day." In the sense I am using it, this student has satisfied the rules of "Formal Dress Day" without following them.

you can satisfy this rule through ignorance and sloth, won't everyone in the world have satisfied it? And perhaps every*thing*?

To deal with this worry contrast Jim with his housemate, Lars. Lars is also ignorant of the fact that the avant-garde art assignment has been given and sleeps through the day of the class when it is due, but this is because Lars is a math major and has not even enrolled in the art class with Jim. This puts Lars outside the scope of the avant-garde art assignment and this is what makes it the case that Jim is eligible to satisfy the assignment by ignoring it, but Lars —and the rest of the universe not enrolled in the art class— are not. This is another source of potential failure, and one that might be doing some work here. Surely, Jim might have failed to enroll in the art class like Lars and thereby failed to satisfy the rule. Could this be the relevant possibility of failure that enables Jim to satisfy the rule in the way he does?

It is probably right to say that, had he not enrolled in the class, Jim would not have satisfied the rule of the avant-garde art assignment, just as Lars did not; but it is a mistake claim that Jim would have thereby been able to *break* the rule. What Lars's case shows us is that there is perhaps a weak sense in which one could fail to satisfy this rule, but only by evading it and managing to stay beyond its scope. Once someone falls within this rule's scope of application it is impossible to break, but one might still fail to satisfy the rule by failing to be subject to the rule at all. Perhaps it is better to say that it is only possible here to do something that is not in accord with the rule, because it is not subject to the rule in the first place. This is a persistent possibility of failure, but it is not one that will vindicate the Wittgensteinian version of the fallibility constraint. That version demands that a rule, in order to be genuine, must have something that counts as *violating*

it, if anything is to be able to accord with it. In avoiding the scope of the assignment, Lars has certainly not managed to either violate or accord with it, because he is simply not subject to it. This way of failing to satisfy the rule by avoiding it does not capture the distinctive sort of violation that the Wittgensteinian claim meant to capture.

Anyone still resistant at this point (those who would want to insist that Lars's case establishes a real and relevant possibility of failure) should note that the example is easily modified so as to have a universal scope. Say, for example, I declare myself to be a mad prophet and order everything in the universe to either bring me a kimono or not. Even though everything in the universe will count as satisfying this demand, that fact will not prevent some of my disciples from doing some things that will count as following it.

Jim's case allows us to see that it is possible to unintentionally satisfy a rule. Jim manages to satisfy the rule without even attempting to follow it, but there is another interesting class of cases where agents can unintentionally satisfy a rule while attempting to engage with it. It turns out that it is possible to unintentionally satisfy a rule both while attempting to follow it and while attempting to flout it. To see this, contrast the cases of another pair of housemates, Deb and Alyssa.[17]

Deb is a clever and dutiful student, but is also prone to accidents. Deb intends to complete the avant-garde art assignment, but, having been distracted by other things, leaves the assignment until just hours before it is due. She has decided to construct a collage of North Korean propaganda posters, along with their most literal translations into English. Scrambling to make it to class on time, with the glue on the collage still drying, Deb races into the classroom, but trips at the last minute and inadvertently tosses

[17] I forget which elements, but I am sure that the following discussion was fruitfully influenced and expanded in discussion with Ted Morris and Charlotte Brown.

her wooden framed collage into the air, only for it to be impaled on tip of the class's flagpole bearing the American flag.[18]

Granting that this is not what Deb meant to do, it should still be obvious that this will count as a perfectly legitimate submission for the avant-garde assignment. "Brilliant!" we can imagine the instructor exclaiming, "I love it! The urgency of your running, the symbolism of the impaling, your theatrical pratfall, everything was wonderful, Deb! Excellent work." Though we can share Deb's knowledge that the instructor has thoroughly misinterpreted her artistic intentions, he has not misinterpreted her inadvertent piece of performance art as a legitimate submission for his assignment. After all, absolutely anything Deb did would succeed in satisfying the rule in that assignment, and in this case she really was following this rule, it just turns out in this case that she did not satisfy it in the way she intended.

But note that the fact that Deb was indeed intending to follow the rule in this case bears no weight in supporting the conclusion that she was able to satisfy and thereby follow this rule. To see this, consider Alyssa.

Deb's housemate, Alyssa, is also enrolled in the art class and takes her education very seriously. This stems from her general seriousness and distaste for the absurd. Immediately on hearing the avant-garde art assignment, she is disgusted. She is enraged that her tuition dollars are being spent to learn such trivial and facile lessons and she sets out to refuse to participate. A savvy student, she knows full well that simply refusing to submit anything could be misinterpreted as mere sloth or artistic attempt. (She knows Jim and Albert.) So, that very evening, she begins a letter writing campaign aimed at both local and national news outlets, expressing her disgust and contempt for this assignment

[18] Thanks to Sean Kelsey for first presenting me with a case like Deb's.

and its frivolity. Her sharp wit and tightly written prose garners a number of admirers, and by the time the day to submit the assignment comes, she has organized a demonstration and picket line against the assignment. As she makes clear (and loud) through her bullhorn, she and her compatriots will not stand for this bumfluffery.

"Fantastic!" shouts the instructor. "This is just a magnificent example of what the avant-garde is all about! Rejecting conventions! Breaking down barriers! Great work."

Despite all her efforts to the contrary, Alyssa has only managed to signal her discontent with this assignment and her intention to refrain from satisfying its demands. But it is no use, as the instructor has correctly (if over-enthusiastically) observed, her letter-writing campaign, demonstration, and declamations collectively count as a perfectly legitimate way to satisfy the avant-garde art assignment. Whatever her intentions here, she has nevertheless managed to satisfy the assignment's rule.

It is an interesting question if Alyssa has followed the rule or not. She clearly understands the assignment, at least well enough to see that it will accept null submissions, and is, after all, pretty silly. She is mistaken in her hopes that her acts will count as breaking the assignment's rule, but she still seems to be deriving guidance from the rule, although she is attempting move in precisely the opposite direction from the one the assignment dictates. I think it ultimately an unimportant semantic question if we are here to count Alyssa's behavior as an instance of *following* or not. What seems clear is that she is deriving guidance from the rule, in that she is trying to do the opposite of what it demands. The sense of "following" that interests me occurs when agents organize their thinking around the rule, even if they greet it with contrarian intentions and some factual errors in their understanding of what the rule demands and will accept. The fact that she

is unable to do anything that counts as refusing its demands does not diminish the fact that she has some understanding of the rule and is being inspired to take a particular course of action as a result. What is important here is to notice is that Alyssa is able to satisfy this rule even while actively intending not to, and so there was nothing in Deb's attempting to satisfy the rule that enabled her to do so.

Alyssa's case also shows us that it is not possible to guarantee compliance to a rule in the sense of universal following simply by guaranteeing compliance in the sense of guaranteeing universal satisfaction. As Alyssa's protests show, it is possible to struggle and make your intentions to refuse evident even against a rule you cannot possibly manage to break.

As I hope the cases of Albert and Ernestine have shown, it is perfectly possible to follow a rule that cannot be broken. As the further cases of Jim, Lars, Deb, and Alyssa should have illustrated, there are still a number of potential possibilities of failure that might surround even an unbreakable rule, but none of them are enough to salvage the Wittgensteinian interpretation of the fallibility constraint that only rules that can be broken can be followed.

III. The Return of Kripkenstein

I have just argued (and am convinced) that the Wittgensteinian analysis of rules and rule following cannot be the basis for a legitimate fallibility constraint in ethics. The Wittgensteinian considerations about rules are not enough to establish a claim along the lines that there is no normativity if you cannot be wrong. Before concluding this

discussion, I would like to flag and guard against two potential objections that Kripkensteinians might bring against this claim of mine. The first is that an adequate solution to the rule following problem will exclude unbreakable rules as the sort of thing that can be followed. The second is that my avant-garde art example is subject to just the same sort of infallibility worries that I brought against Divine Command Theory above. I will now address these objections in turn.

i) Kripke provides an extensive discussion of Wittgenstein's Remark 201 in the second chapter of *Wittgenstein on Rules and Private Language.* Largely, Kripke's considerations here are tangential to my current considerations with my arguments above. Kripke is concerned with elucidating the difficulty with rule following in general. As Kripke reads him, Wittgenstein is making the metaphysical suggestion that there might just be no such thing as following a rule, not just the epistemic difficulty of ever discerning which rule is being followed. By contrast, my concerns above have been to establish the claim that provided that there is a way to follow a rule, an unbreakable rule can be followed just as much as a breakable one can. So far Kripkenstein and I are speaking past each other.

But in his third chapter, Kripke gives his reading of Wittgenstein's solution to the rule following problem that Kripke went to great pains to explain. It is here that a Kripkensteinian might be making claims that could be in conflict with my claims about the normativity of unbreakable rules. This is because the solution Kripkenstein offers to the rule following problem looks like it might exclude unbreakable rules as the sort that could be followed.

This is because the solution depends on shifting our focus from a piece of mental content in the mind of the rule follower to a community's overall practice of rule following. Because there is no fact about my present mental states that commits me to any particular future interpretation of my words, I must instead appeal to the practices of a community. Whatever my mental content relating to a particular word is at any given time, it is consistent with any number rules of interpretation for that word. It is only in the context of a practice of rule following that we could generate anything like the following of a rule. In order to do this, "the community must be able to judge whether an individual is indeed following a given rule in particular applications, i.e. whether his responses agree with their own." (WRPL 109)

If this is what it takes to generate even a "sceptical solution" to the rule following problem, this might make it the case that breakable rules can be followed, but unbreakable ones cannot. This could be because, if all we can manage of rule following involves a community being able to correct an individual when he makes a mistake, perhaps this means that a community can only correct an individual if there are things that the community would count as inconsistent with their practice. The worry is that if the rule governing the community was something like "Every use is correct," then however much judging the community might do, they would never find an instance of a mistake, and so heeding their judgments would be no different from simply ignoring them. But then what function is the appeal to the community serving? How can it provide a solution in this case if it is functionally irrelevant? If the only workable solution to the rule following problem cannot work for unbreakable rules, this will be a serious problem for my claims in this chapter.

If a community attempted to establish an unbreakable rule, they could never correct one another, and so there could be no rule following in such a case. Therefore, the Kripkensteinian might argue, Wittgensteinian considerations can establish a fallibility constraint for ethics, because the only solution to the rule following problem shows that unbreakable rules cannot be followed in the only real sense that a rule can be.

I have put this argument in the subjunctive mood, because I think it is weak enough that I would not want to saddle the Kripkensteinian position as automatically committed to it. This is because I take it that the Kripkensteinians and I substantially agree that what is important about rule following is not so much a matter of content as it is a matter of practices. For those following Kripkenstein, the point is to shift our focus away from mental contents. For me, the point is to shift our focus away from the content of rules.

In the previous section, all of my descriptions of students responding to the avant-garde art assignment just *are* descriptions of a community building a practice around an unbreakable rule and checking one another against it. When I invite us to take the instructor's perspective on the students' various submissions, what I am doing is inviting us into the community with the students and instructor so that we can discover that, yes, those submissions are such that we could imagine ourselves giving as appropriate responses to the rule governing the avant-garde art assignment. Simply because, as it happens, this community will never correctly find someone to be in violation of the rule its practices create does not mean that the members of the community cannot judge one another to be behaving appropriately. My very act of arguing to my readers that these

practices do count as appropriate responses to the given rule begins to constitute the sort of community that the Kripkensteinians think make rule following possible.

ii) Even though I think Kripkensteinians should not disagree with me on the issue outlined above, they might still be inclined to object that I have failed to properly engage their position. As I myself noted above, their paradigm examples of troublesome unbreakable rules have the form "whatever seems right, is right." So, if I were to have my avant-garde art assignment be for the students to turn in "whatever seems right," I would then be able to see how this rule could be neither broken nor followed and *that* is how Wittgenstein's considerations about the nature of rules could establish a fallibility constraint that could stick.

There are at least two problems with this line of objection. First, it seems that my students could *follow* this rule perfectly well, it is only that the instructor might not be able to evaluate if the students had successfully *satisfied* it or not. But this would just be an epistemic problem based on the limitations of the instructor's epistemic access to his student's judgments about what seems right, not the kind of deep metaphysical problem Kripke reads Wittgenstein as establishing.

Far more troublingly, though, is the fact that the rule "whatever seems right is right," is not even an example of an unbreakable rule. In order to get it to be unbreakable, one has to add the substantive view in moral psychology that agents can never against their own best judgment.

Perhaps Thomas Hobbes proposed a moral psychology according to which this would be true. As he says in *Leviathan*,

> In deliberation the last appetite or aversion immediately adhering to the action or the omission thereof, is that we call the WILL ... *Will* therefore *is the last appetite in deliberating*.[19]

Given this sort of rudimentary moral psychology it might well be the case that agents can never do other than seems best to them, since their will is just whatever appetite immediately precedes their acting. But Hobbes' moral psychology is far from the obviously correct description of the mental life moral agents such as we are experience. Anyone who admits a phenomenon like weakness of will has already conceded that it is possible to do other than what seems best. To suffer from weakness of will is just to end up doing other than what seems best.[20]

Whether or not a Hobbesian moral psychology is the correct description of our moral psychology, the point is that a rule to the effect that "whatever seems right is right, so do what seems right" is *not* obviously unbreakable. On any moral psychology on which it is possible to act contrary to one's own best judgment, it will obviously be possible to break this rule. Even in the most banal sort of weakness of the will case, it seems best that I not eat the chocolate cake, yet I find myself eating it nonetheless. So much for the claim no one could violate the rule "do only what seems best."[21]

This is why, as I mentioned in a previous footnote, I am not convinced by the fallibility based argument I laid out against Divine Command Theory. It is pitched against a moral theory in just the way the fallibility constraint should be, but it cannot be a successful criticism of Divine Command Theory because the rule it proposes is infallible can be broken.

---

[19] Hobbes, 1994. Part I, Chapter vi, paragraph 53.
[20] For other interesting and compelling cases of acting against one's best judgment, see Arpaly, 2003.
[21] Insisting that this misses the force of the linguistic nature of this rule will not help. People recognize themselves as misspeaking all the time, and discover that they have used (perhaps accidentally) a word in a way that did not seem best to them. Anyone who has ever been flustered has experienced this.

IV. Unthinkable Actions and Deliberative Paralysis

In closing, I would like to address one potentially persisting worry and leave those who would disagree with my overall thesis in this chapter with a challenge.

This is the nagging worry that there is just something deeply wrong with the idea that anyone could count as *following* a rule that genuinely leaves open absolutely every potential course of action. If, to follow a rule, one must choose some course of action, there is a worry that genuinely unlimited options would *prevent* an agent from choosing in the first place.

This is what Harry Frankfurt thinks in his "Rationality and the unthinkable."[22] There, Frankfurt lays out just this sort of problem, and is worth quoting at length.

> Now suppose that the field of alternatives from which a person may select is not merely extended; suppose that its boundaries are wiped out entirely. In other words, suppose that now every possible course of action is available and eligible for choice, including the course of action that would affect the person's preferences themselves. Since he can in that case even alter his own will, it seems that he has to confront the choices he must make without any specific volitional character that is definitively his. The person's will, we are supposing, is whatever he chooses it to be. Therefore, it is nothing until he has decided what will to choose.
>
> But how, then, is he to make any choice at all? What preferences and priorities are to guide him in choosing, when his own preferences and priorities are among the very things that he must choose. ... He may possibly remain capable of some hollow semblance of choice. If he does, however, it will only be by virtue of a vestigial susceptibility to inchoate volitional spasms. (RU 177-178)

If Frankfurt is right that having every possible course of action will "doom" us "to such paralyzing volitional emptiness," then perhaps following an unbreakable rule will be impossible after all, because it creates a context in which any choice among options is

[22] Frankfurt, 1988b. Hereafter referred to parenthetically in the text as RU with page number(s).

impossible. The only way out is when one is arbitrarily bumped out of this deliberative paralysis by an "inchoate volitional spasm." (RU 178)

Frankfurt's own solution to this problem, or at least explanation as to why we are not chronically suffering from it, is to point to what he calls "unthinkable actions." These actions are those that an agent cannot bring himself to do, even once he decides to carry them out and so serve as "anchors" in his will, protecting it against deliberative paralysis by permanently removing some options from deliberative availability, and thereby providing settled points from which to choose and reason further.[23]

The way we avoid the deliberative paralysis that would set in if absolutely every choice were available is by finding a way to exclude some options. Otherwise, we might never be able to choose at all.

The prospect of an unbreakable rule seems as if it would create a context wherein all options were available and equally attractive, threatening deliberative paralysis. But this is to misunderstand the nature of an unbreakable rule. What unbreakable rules do is make every course of action equally *permissible* by their own lights; they do not therefore make every course of action equally *choiceworthy* by any given agent's lights. Agents can still choose among these equally permissible courses of action on the grounds of some other considerations that matter to them, just as the characters I have described above do. According to the unbreakable rules I have given, all courses of action are in a

[23] As it happens, I think Frankfurt has described the wrong sort of entity to solve the sort of deliberative paralysis he describes, but has given this entity just the right name. As Frankfurt describes them, "unthinkable" actions are still available in deliberation. An agent only discovers that a given action is unthinkable when he settles on one as a course of action, only to discover that he cannot carry it out. Such an action is not, "unthinkable" only "undoable" and so cannot possibly solve an essentially deliberative problem where the difficulty is which course of action should be settled on. What might be able to fix this problem would be a course of action that is more literally "unthinkable," in the sense that the agent could not even consider carrying it out. I hope to develop this point further in a future paper.

universal tie as equally permissible, but this does not require that an agent must find them all equally attractive.

Along the lines of deliberative paralysis, it might be tempting to object that the rule I have proposed in the avant-garde art example cannot count as genuinely normative because it does not recommend a unique course of action. This is not a complaint that can be made to stick, though, for it rules out some of the most obvious cases of normative requirements. Underdetermination and plumping from indifference are not new wrinkles in normative theories. Utilitarians have long been aware that the utility calculus might reveal multiple courses of action that are all equally maximal. Any Kantian who recognizes imperfect duties will also face a similar problem, since imperfect duties give agents significant latitude in how, when, and in what way to satisfy them. Even intuitive moral rules that incidentally count as perfect duties for Kantians can be radically underdetermined. The injunction to never lie leaves open as permissible every course of action that does not consist in lying.

Take a rule as simple as "Don't smoke." It gives me enormous latitude in how I shall follow it, but this does not mean I am actually following my further considerations and not the rule. Suppose I decide to obey this rule by going to the theatre, where I cannot buy cigarettes and will be found out and punished if should smoke any I might sneak in with me. Even though there are many other ways I might have chosen to obey this rule, my deciding to follow it according to what I think will best restrict me from smoking (given my proclivities) hardly shows that I am therefore not following the rule, because I settled on a particular course of action based on criteria not specified by the rule itself.

If there is a problem with the cases I have constructed it cannot be in the fact that the rules do not recommend a unique course of action. This is a point worth stressing, because it allows me to shift the burden of proof. If there is nothing generally wrong with having a rule that can be satisfied in multiple ways, then it seems that adding one more way of satisfying it cannot eradicate its normativity.

For example, when Ernestine submits her sculpture to fulfill the assignment that will accept all but null submissions, she has a huge number of ways of completing the assignment.[24] When she submits the same sculpture to satisfy an assignment that will also accept null submissions, she has just one more way to complete her assignment, but she still has a rule she can follow and against which she can be evaluated. Those who wish to establish that unbreakable rules cannot be normative must be able to explain why the rule in the first case is normative while the rule in the second case is not.

Some, and I count myself among them, might not even be tempted to distinguish the rules in the two cases, because the considerations above seem enough to settle the question. They will conclude that Wittgensteinian considerations about rules cannot be the ground of the fallibility constraint. Some in this group, and I count myself among them as well, might find this conclusion alarming because they have commitments, and perhaps even further arguments, that depend critically on the truth and justifiability of the fallibility constraint. In the following chapter I hope ultimately to quiet this alarm, by suggesting where a stable justification of an ethical fallibility constraint can be found. It should be unsurprising that the proper place to look is around the question of moral

[24] In fact, Ernestine has an infinite number of options. She could have turned in a painting of a woman with one nose, or a painting of a woman with two noses, or a painting of a woman with three noses, etc.

responsibility, but it is surprising where in moral responsibility a justification can be found.

Chapter 2

Moral Responsibility and the Possibilities of Failure

Just because the fallibility constraint cannot be justified by appeal to Kripkensteinian considerations about the nature of rules, this does not mean that it is yet time to abandon all hope that some kind of fallibility constraint can be justified and so the foundation for the rest of the philosophical work built around it can remain stable. In this chapter, I aim to show that such a justification for a fallibility constraint in ethics can be found by examining moral responsibility. Concentrating on alternate possibilities might seem like the obvious way to garner such a justification, but as I will argue below, the possibility of doing otherwise in no way guarantees the possibility of failure. Instead, I will claim that we have strong independent evidence to believe in ill will, the failed instance of the kind whose possibility we seek to establish here. Belief in actual ill will is obviously more than enough to establish its possibility, but merely accepting its coherence is sufficient to establish this possibility as well. Those who would wish to deny this possibility will have to pay a heavy theoretical price by way of losing the reactive attitude of resentment and its apparent fittingness as a response to ill will.

But before I detail all of this, I would like to begin this chapter with a brief introduction pointing out the deep intuitive connection between moral responsibility and at least one kind of possible failing.

I. The Intuitive Connection

It is hardly an innovation to see moral responsibility as resting, at least partially, on the possibility that an outcome might fail to obtain. If an outcome's occurrence is guaranteed by some force beyond an agent's power, then it seems inappropriate to hold that agent morally responsible for that outcome, precisely because the agent was unable to prevent it. Here, the inaccessible possibility of failure that seems relevant is in the fact that the agent could not make that outcome fail to obtain. If the world is at no risk of failing to be a particular way, it seems that the way the world is can only be occasion to rejoice or lament, not to hold responsible.

Ideas like this are at least as old as Aristotle. In the *Nicomachean Ethics*,[25] he notes that

> Since virtue is concerned with passions and action, and on voluntary passions and actions praise and blame are bestowed, on those that are involuntary pardon, and sometimes also pity, to distinguish the voluntary and the involuntary is presumably necessary for those who are studying the nature of virtue, and useful also for legislators with a view to the assigning both of honours and punishments.
>
> Those things, then, are thought involuntary, which take place under compulsion or owing to ignorance; and that is compulsory of which the moving principle is outside, being a principle in which nothing is contributed by the person who is acting or is feeling the passion, e.g. if he were to be carried somewhere by a wind, or by men who had him in their power. (NE 1109b30-1110a5)

Though Aristotle does not make alternative possibilities the centerpiece of his observation here, the idea is obviously at play. In his two examples of involuntary (passions and) actions, the agent's capacity to resist either the wind or his captors is presumably closed off. If the agent had the capacity[26] to resist and merely allowed

[25] Aristotle, 1941. Hereafter referred to parenthetically in the text as NE with citations to the Bekker numbers.

[26] For Aristotle, this capacity will include the knowledge of both the power and the particulars of its use.

himself to be carried along, this would make occurrence voluntary, since the person contributed to its happening by declining to prevent it.[27] Since it was up to the wind or the captors what happened, the agent is unable to cause or prevent any given outcome. If the outcome is bad, it seems we ought not hold the agent responsible for that bad outcome precisely because the agent could not make that outcome fail to obtain.

As an agent's capacity to influence alternative outcomes changes, so too do our intuitive judgments of moral responsibility.[28] Observing some of these apparently interdependent changes is enough to indicate a connection between moral responsibility and this possibility of failure. If a wind were to pick me up and carry me through your window, it would be madness hold me responsible for the window's breaking. For, once the wind was powerful enough to lift me off the ground, I could no longer fail to move in any way it directed me.[29] While this fact seems to work perfectly well to absolve a flightless human such as myself from any moral responsibility for breaking your window, a similar excuse would simply not be available for Superman.

Since describing things in terms of "failure" generally carries negative connotations it feels a little unnatural to couch acts with positive moral value[30] in this way, but it is important to note that this possibility of failure (i.e. the possibility that an

---

[27] Though it is still somewhat controversial, I am fully on board with the idea of actions by omission. For more on this see Rachels, 1975.

[28] What counts as capacity is difficult and controversial. Since I am here only warming up the intuitions that moral responsibility involves the capacity to fail, I will not endeavor to sharply delineate the logical space here. For an excellent discussion of the various kinds of capacity, see Watson, 2002.

[29] Obviously, the sense in which I could not fail to move as the wind directed me is not matter of deep logical or conceptual necessity, it is only to observe that once the wind separated me from any way to push against it, it is no longer *within my power* to make myself fail to land where the wind directs me.

[30] I am being deliberately vague here in my description of acts having either "positive" or "negative" moral value. This is to both bracket and remain neutral among the interesting sets of questions about in what moral value consists and where the lines ought to be drawn among the actions that warrant praise, blame, or something else entirely. The ambiguity of "positive moral worth" is especially useful in allowing me to remain silent on the issue of the difference between obligatory and supererogatory moral successes. For illumination of this distinction, I recommend Vanessa Carbonell's work.

outcome might have failed to obtain) applies as much to the moral responsibility for the good actions as the bad.[31] The connection seems to be between moral responsibility and the possibility that a given outcome might have failed to occur, whether the outcome was good or bad, whether the act was virtuous or vicious.

The strength of this intuitive connection is often thought to be captured in if not simply stated by the Principle of Alternative Possibilities, and so it is a sensible place to look first for the grounding of a kind of fallibility constraint on moral responsibility. To clarify, these observations are simply to ostend to the common idea that moral responsibility is tied up with some kind of failure; so far I wish to make no commitments as to what kind. I merely wish to convince the reader that moral responsibility is a promising next site to search for a justification of some sort of fallibility constraint. It will turn out not to be the right place to look, but to see why, we must examine it all the same. It is to this examination I now turn in the next section.

II. Fallibility and Alternate Possibilities

i) The Principle of Alternate Possibilities (or "PAP," as it has come, regrettably, to be abbreviated) insists "that a person is morally responsible for what she has done only if she could have done otherwise."[32] If this is correct, then it might well justify the fallibility constraint because it appears to just *be* the fallibility constraint. It simply states that there can be no moral responsibility if a person could not have failed to do as she did. Those incompatibilists who subscribe to the Principle of Alternative Possibilities will

[31] For a brilliant case for the claim that this intuitive symmetry breaks down see Susan Wolf, "Asymmetrical Freedom," (Wolf, 1980).
[32] Widerker and McKenna, 2006. Preface.

probably have thought all along that there is something like a justified fallibility constraint in moral philosophy, since their position is automatically committed to one. These incompatibilists might seem to be my most natural allies in securing a fallibility constraint in moral philosophy, given that their position looks to entail one. If the relevant possibility of failure that can vindicate the fallibility constraint is just the abilty to fail to do as one did, i.e. the ability to do otherwise, then those who subscribe to PAP will already be committed to a kind of fallibility constraint. But as I will argue, while the Principle of Alternate possibilities does involve something that could sensibly be called a "fallibility constraint," it cannot ground the sort of fallibility constraint that moral philosophers use to show that a proposed moral theory cannot be normative.

This is signaled by a rhetorical clue in the dialectic surrounding considerations of the fallibility constraint. Thanks to the list that Douglas Lavin has already compiled,[33] we can see that when philosophers do speak about the fallibility constraint, they treat it as if it were an *a priori* truth, knowable by simple reflection and the sort of thing about which there could not be reasonable disagreement. (This is surely part of why Wittgenstein's epigrammatic statement about rules seems to be an elucidation of the fallibility constraint.)

While the posture that thinking otherwise is impossible is one that philosophers sometimes adopt toward genuinely controversial doctrines, the fallibility constraint is never treated as a claim about which there is even confused disagreement. Certainly, it is legitimate for most philosophers not to acknowledge the fallibility constraint as a contested doctrine, because it is only very recently that the fallibility constraint (or "error constraint" as Lavin calls it) has even come into question in terms of its proper statement

[33] See Lavin, 2004. pp. 424-425.

and justification.[34] This standard treatment of the fallibility constraint as obviously and uncontroversially true immediately distinguishes it from its potential ground, the Principle of Alternate Possibilities.

Ever since the cases from Frankfurt's 1969 paper "Alternate possibilities and moral responsibility," gained fame, the Principle of Alternate Possibilities has ceased to be the sort of claim that could be seriously treated as obviously or uncontestably true.[35] Some might take the threat that Frankfurt's cases pose to the PAP to be similarly menacing to any fallibility constraint in ethics, but far from doing this, Frankfurt's cases provide the first clue that a legitimate fallibility constraint in ethics cannot be derived from or reduced to the Principle of Alternate Possibilities. Frankfurt's cases have done this by rattling the foundations of PAP in the eyes of many philosophers, while leaving the fallibility constraint untouched. This is a telling clue.

Surely, clues can mislead, but this one does not. A little further reflection will reveal that the Principle of Alternative Possibilities is not even capable of justifying the fallibility constraint as it is employed by the philosophers who invoke it explicitly. Even if PAP is correct and grounds a *kind* of fallibility constraint, it cannot be one that can support the philosophical work that rests on what I have been calling the fallibility constraint.

To see that PAP's fallibility constraint cannott do the trick, let us consider an example. Perhaps the paradigm example of the fallibility constraint in action is Korsgaard's use of it against Humean accounts of practical reason in "The Normativity of Instrumental Reason." Her complaint there is that Humean accounts of practical reason

[34] Though many philosophers use it, and briefly state or allude to it, Douglas Lavin's 2004 paper was, to my knowledge, the first effort to state and examine the "error constraint" to make it to print.
[35] Frankfurt, 1988a.

cannot be normative because they always conclude that what an agent does is in fact what that agent has most reason to do, and so an agent, on this picture of the Humean account, can never be practically irrational. As Korsgaard herself puts it,

> Hume identifies a person's *end* as what he *wants most*, and the criterion of what the person wants most appears to be what he actually *does*. The person's ends are taken to be revealed in his conduct. (NIR 230)[36]

If this is an accurate picture of Hume's account, then it will indeed mean that Hume's view will not permit agents to do otherwise than be rational. This is not, however, the sense of not being able to do otherwise upon which the Principle of Alternative Possibilities insists.

Korsgaard herself is aware of this when she says that "Hume's view is not just that people don't *in fact* ever violate the instrumental principle. He is actually committed to the view that people *cannot* violate it." (NIR 228) The instrumental principle recommends that agents take the means to their ends. As Korsgaard reads Hume, this will mean that an agent's ends are "revealed in his conduct" because the conduct the agent is displaying indicates the end he has chosen. When Hume says that, "'Tis not contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter,"[37] he claims that there is no *irrationality* in my seeking even what I would claim is not my greatest good.

There are two types of behavior that earn the title "irrational" on Hume's view. These are instances in which one's practical reason is misled by being based on a non-existent object or a false judgment of a causal relation.

---

[36] I should be careful to note here that this does not mean that an agent cannot fail to have false beliefs or cannot fail to have successful attempts. All it means is that an agent's actions will reveal, in light of that agent's understanding, what that agent wants most. Thanks to Victor Caston for pushing me on this point.
[37] Hume, 1978. Treatise 416, Hereafter referred to parenthetically in the text at T with page number(s).

Korsgaard is well aware of these two types of Humean irrationality and describes them as "not, strictly speaking, irrational." (NIR 228) This is likely because they involve no *practical* irrationality, only theoretical mistakes about what exists and how things are related. These are just false beliefs, and the error comes from reason treating them as if they were true. In this case it is not really *reasoning* that is misfiring, it is belief.

Therefore, whatever I do is rational for me to do on this picture, because my so acting both determines what my strongest desires are and reveals what they are. I might lust after an object I do not yet know is only imaginary, or seek an object in a backward way, but I do and can never fail to pursue my strongest desire since my very pursuit makes it the thing I want most. Types of irrationality like weakness of will are just not possible on this understanding of the Humean picture.

On Korsgaard's picture of Hume, this is how is always has to be for every agent who might appear to other accounts as irrational. Whatever they might report about their strongest desires, it is their actions that settle the question of what they want most. Thus, when other accounts might be tempted to declare me irrational or weak willed when I assiduously avoid a wedding that I declare to be my heart's desire, this version of the Humean account will only find me guilty of a minor theoretical miscalculation about what my desires really are. That, or a lie.

ii) With this clearer picture of the sort of infallibility Korsgaard finds in Hume's account, we can see that it does not, and cannot, reduce to a complaint deriving from the Principle of Alternate Possibilities.

Agents cannot violate this understanding of the instrumental principle not because their hands are perpetually forced by a deterministic universe so as to always obey it, but rather because there is no behavior that *counts* as violating it. It *is* metaphysically impossible to be irrational on this picture, but only because it is conceptually impossible to behave irrationally according to its definitions. This picture of Humean practical reason has it that whatever a person does is just what the instrumental principle recommends that the agent do. Since one's conduct settles the question of what one's end is, there simply is nothing that an agent can do that could count as not taking the means to her end, since the means she takes settles which end it is she is and ought to be pursuing.

This sort of infallibility not only does not reduce to the Principle of Alternate Possibilities, it can remain just as vexing even when that Principle is satisfied. Since they can be satisfied independently, they cannot reduce to one another.

First, PAP can be satisfied even when this sort of infallibility persists. Grant an agent the fullest libertarian freedom required to satisfy even the most ardent advocate of the Principle of Alternative Possibilities, and Korsgaard's complaint about the Humean theory of practical reason can still be made to stick. Whatever this agent chooses with his libertarian freedom will be just what he should have chosen according to this reading of the instrumental principle. Since in choosing a particular means, the agent thereby makes it the case that she most wants the end those means lead to, *anything* she does will count as practically rational on this picture.[38] Even if this agent was fully metaphysically free and able to realize any number of alternative possibilities, this Humean picture of

[38] Of course, there could always still be instances of irrationality deriving from non-existent objects and false causal judgments, but, as we have seen above, these are not the distinctively practical failings that Korsgaard wishes to characterize.

practical reason will still declare the one that was realized to be the one that should, rationally, have been realized. It is this sort of infallibility Korsgaard wishes to identify, and that can remain even when the Principle of Alternative Possibilities is satisfied.

Further, the sort of fallibility that Korsgaard takes to be necessary for normativity can be present even when the Principle of Alternative Possibilities is clearly not satisfied. Suppose another practical theory had it such that the rational actions were only those that resulted in overall profit. Placing this theory in a universe with even the most pernicious sort of determinism will not prevent it from providing a kind of fallibility that would indicate normativity for Korsgaard. If the inexorable chain of determined causes push some agents into investments that fail to aggregate to profit, there will be obvious proof that it is possible to fail to be practically rational on this picture even when while the Principle of Alternative Possibilities is not satisfiable. Even if all agents ended up being determined to do only that which is profitable, the distinctively conceptual sort of fallibility is still available, because there would still be something that *counts* as being irrational, even if it is never realized.

The kind of infallibility that worries Korsgaard here is quite indifferent to *how* an agent comes to perform any particular action, because her complaint is precisely that *however* and *whatever* an agent does, this Humean account will declare it to be practically rational, and so there can be no practical irrationality by its lights. The problem Korsgaard wishes to identify in Hume's account rises out of the conceptual particulars of Humean practical reason, not the metaphysical particulars of freedom, and so the Principle of Alternative Possibilities cannot be the proper ground for the sort of fallibility constraint that Korsgaard is employing here.

Once these two distinctive types of failure are clearly separated from each other it should be especially obvious that the sort of fallibility that worries Korsgaard could never have been the sort of fallibility that the PAP insists upon, since they are of different modalities. One concerns the conceptual possibility of counting as wrong, while the other concerns the metaphysical possibility of doing otherwise. Both could legitimately be described as involving a "possibility of failure," but involve not only different sorts of failings, but different sorts of possibility as well.[39]

If this sort of conceptually guaranteed infallibility truly is a fatal error for a normative theory, it cannot be because of the grip the Principle of Alternate Possibilities has on moral philosophy. The metaphysical constraints it puts on this kind of moral responsibility cannot be the source of this other conceptual fallibility constraint.

This fact does undermine the most immediately obvious reason why an incompatibilist who endorses the Principle of Alternative Possibilities might be my ally in seeking to ground a fallibility constraint in ethics, but this hardly means that there is no reason for such an incompatibilist to agree with me. There is still a reason I can offer to the incompatibilist, and this reason has an added benefit. The benefit is that it will allow a compatibilist who rejects the Principle of Alternative Possibilities to also accept the conceptual fallibility constraint and my justification of it. This is because, as I hope to show in the next section, there is a vindication of the fallibility constraint that is compatible, more or less, with the views of compatibilists and incompatibilists alike.

Before I continue, allow me to concede that my proposed justification will not be amenable to all sorts of incompatibilists. The incompatibilism that finds free will to be incompatible with a deterministic and materialistic universe because such a universe will

[39] My thanks to Ali Kazmi for putting the point to me this way.

exclude all mental content, cannot accept a solution based on ill will and resentment. They are both mental entities, the attitudes of beings capable of forming mental representations about both how the world is and what its value is. The incompatibilist who finds all such content impossible if determinism is true can find no consolation in my solution proposed below. This is not a problem that vexes me, however, since even though I may lose a philosophical ally here, I do not lose him to another competing camp in this debate around moral philosophy and the possibility of failure. Such a skeptic about all mental content should be hostile to all accounts of morality and moral responsibility, and so no special problem for my account.

III. An Alternative to Alternative Possibilities

i) This justification is one that can be made to appeal to most compatibilists and incompatibilists alike in part because it derives from a paper that aimed to get beyond the debate between those who thought free will and moral responsibility could be reconciled with determinism and those who could not. This paper is P. F. Strawson's "Freedom and Resentment."[40] There, Strawson famously draws our attention to the importance of intentions and reactive attitudes in our practices of moral responsibility. Attitudes like malevolence and the reactive attitudes of resentment or indignation that are appropriately felt in response to them are part of the "complicated web of attitudes and feelings which form an essential part of the moral life as we know it." (FR 23) In a Humean spirit, Strawson insists that it is a practice we could not manage to abandon even if we discovered it was somehow illegitimate, but, more importantly for present purposes, that

[40] Strawson, 1974. Hereafter referred to parenthetically in the text as FR with page number(s).

neither the truth nor the falsity of determinism could succeed in showing that the practice was illegitimate.[41] Reflecting on Strawson's argument for this will show that there is a deep and indispensible feature of moral responsibility that presupposes the possibility of failure. This feature is "the general framework of attitudes [which] itself is something we are given with the fact of human society. As a whole, it neither calls for, nor permits, an external 'rational' justification." (FR 23)

Strawson's paper was written in 1962 and, for all its influence, is already bearing marks of its age. It was a time when Strawson could remark that "It is a pity that talk of the moral sentiments has fallen out of favour. The phrase would be quite a good name for that network of human attitudes" that Strawson describes. (FR 24) It also involves discussion of something called the "objective attitude" which might confuse the modern ear. This objective attitude is not to be contrasted with a subjective one, but rather the "participant attitude" that we might alternatively adopt when dealing with other humans. The distinction between these two attitudes is difficult to draw precisely, partially because we can adopt both attitudes toward a single entity at the same time, but it is clear that we only adopt the participant attitude toward those people we take to be participants with us in interpersonal relationships.

The objective attitude earns its somewhat misleading name by being objective in two senses. It is the attitude that we adopt toward entities from which we have greater personal distance, and as such can view more "objectively" in the sense of less passionately, and it is the attitude that we adopt toward entities that we take to be less than full participants with us and so consider to be more of a mere "object" as opposed to a fully morally responsible participant. As Strawson puts it, "If your attitude towards

[41] Provided, of course, that the truth of determinism does not exclude all mental content from the universe.

someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him." (FR 9) Participants are those you might negotiate obstacles with; objects are those obstacles you must negotiate around.

Put in more modern terms, participants are those with whom we occupy a shared space of reasons and to whom we can be mutually comprehensible. An agent who should instead be understood in terms of "the purely objective view" is one "whose picture of the world is an insane delusion" or "whose behaviour...is unintelligible to us, perhaps even to him." (FR 16-17) Such a creature might exhibit hostility toward me, even hurling apparent insults at me, but it does not make sense for me to resent his insult, however much it might wound me, in the way it does for me to feel indignation beyond the wound when a participant levels the same insult against me, in a deliberate effort to wound me.

This distinctive cluster of reactive attitudes including resentment, indignation, and gratitude are only appropriate when they come from someone who is a fellow participant with me in the moral community. They are not appropriate reactions to agents I understand to be less than full participants with me in the moral community, even when the insults are precisely the same. Though the dog's bite might hurt more than the stranger's shove, it is appropriate for me to resent the stranger's behavior in a way I cannot resent the dog's. In fact, insofar as I do resent the dog for biting me, I conceive of him as being a participant with me and so be directing a distinctly interpersonal sort of ill will toward me.

What is critical to notice is that the appropriateness of these reactive attitudes depends entirely on beliefs and facts about the understanding and intentions of the agents

who provided the attitude provoking the reactive attitude, and do not depend at all on whether or not those agents were metaphysically determined to do what they did. As Strawson puts it,

> [I]t is not a consequence of any general thesis of determinism which might be true that nobody knows what he's doing or that everybody's behaviour is unintelligible in terms of conscious purposes or that everybody lives in a world of delusion or that nobody has a moral sense... (FR 18)[42]

What can make my resentment appropriate is the fact that another agent's actions genuinely reflect some malevolence that agent bears toward me, not the causal particulars of how that agent came to bear that malevolence against me.

Here it is important to distinguish between two sorts of appropriateness. The appropriateness that determines which reactive attitude is the appropriate pairing with the attitude to which it reacts is simply an issue of fit, and is nearly independent of causal histories. With the exception of indexicals, whose content is determined by their histories, issues of fit depend on content irrespective of how that content came to be. This is distinct from the appropriateness that would warrant actually feeling or expressing a reactive attitude like resentment. Warrant can be governed by mitigating factors, including causal histories. A feeling or expression of resentment might be morally unwarranted if the malevolence is forgivable. It might be epistemically unwarranted if the evidence for the malevolence is poor. It might even be pragmatically unwarranted if it is dangerous to express. But as long as it is a response to the malevolence of a participant agent, it always fits.

This distinction can be used to explain away a potential objection provided to me by Elizabeth Anderson, and explaining the objection away should further illuminate the

[42] To be fair, it is the consequence of the kind of determinism that banishes all mental content from the world, but as I have noted, this will unseat all moral philosophy, not just my downstream considerations.

distinction itself. Anderson suggests we consider particular family members who, as we observe over a number of years, carry around a sort of chronic ill will that frequently gets directed at people but it seems mistaken to take personally and unwise to resent. It seems mistaken to resent it because since we understand them to just be chronically angry, it is difficult to understand their anger as really *about* us, even when it is directed at us. It seems unwise to resent it, because resenting it will likely only cause them to be further irritated and direct more anger at us. Not only does expressing this resentment seem potentially counterproductive, even feeling it seems unwise, since it will likely only bring frustration in the inability to express it or futility in expressing it.

There are two ways of understanding this case, neither of which threatens the fittingness of resentment to ill will. If the anger truly isn't *about* us, say it started long before we arrived, and we just so happened as to get in it way, then it is not fitting to resent, because there are two conditions of fit for resentment: ill will and intentionality. If we merely catch some ill will that is not truly *about* us, then it is not fitting to resent.

Alternatively, if we take the agent to be expressing a kind of anger that is fully a part of his agency and he means ill when he directs it, then the distinctive sort of moral resentment is still fitting. Noticing, however, that it is unwise to actually give oneself over to feelings and expressions of this resentment does not make it any less *fitting*; it may just make it less emotionally or pragmatically *warranted*. Since it is genuinely intentional ill will that is genuinely directed at us, resentment fits the occasion even if the occasion does not call for it.

The metaphysical history of any given piece of malevolence against me might still bear on some further questions of credit, blame, or warrant but it does not impact the fact that resentment is the fitting reactive attitude in response to it.

This is a subtle point that might be unfamiliar and is worth pausing to illuminate. Consider it, therefore, in an analogy to a more theoretical case.[43] On Strawson's view, an attitude fits a reactive attitude in virtue of its content in just the same way a belief is true or false in virtue of its content. Again, excepting indexicals, the content of beliefs and attitudes are separable from the facts about how those beliefs and attitudes came to have their content, and whether the content could have been otherwise. How much credit an agent deserves for holding these beliefs or having these attitudes might depend on these facts, but those facts do not impact what reactive attitude is fitting any more than they settle whether or not a belief is true.

First consider the case of belief. Suppose our epistemic agent is in a room containing a cube, a cone, and a pyramid. Suppose further that he believes that he is in a room containing a cube, a cone, and a pyramid because he has consulted the sense data of the homunculus who lives in his mind and enables him to perceive. Now allow that the world in which he came to have these beliefs was one in which the most pernicious sort of determinism was true, and that he could not have done otherwise than to believe that he was in a room containing a cube, a cone, and a pyramid. Whatever implications this might have for the amount of credit he deserves for holding these beliefs or the amount of trust we should put in him as an epistemic agent, his beliefs are nevertheless *true*. Even if we complain that, beyond the determinism, this agent formed his beliefs by a method that compounds terrible theories of perception and so should not be trusted (if such

[43] Thanks to Peter Railton for suggesting this analogy.

methods are even possible), the content of his beliefs still accurately represents the world and so they are still true. Truth depends on content, not on origins.

Compare this with the case of a reactive attitude. Suppose our practical agent comes into his office one day just as one of his officemates, fed up with his affectation, is smashing his favorite ornate teacup. Allow that the officemate who is smashing the teacup truly does bear ill will toward our practical agent and is smashing the teacup as a way of expressing this ill will. Given these facts, it might be *warranted* for our practical agent to feel or express his resentment, but even if it was not, that resentment would still be fitting.

Now suppose that this scenario also occurred in a world where determinism was true, and the officemate could not have done otherwise than to smash the cup. Here, incompatibilists will insist that the officemate can deserve no blame for doing so, while compatibilists will insist that there might still be a way that the officemate can deserve blame for doing even what he was causally determined to do. This issue of blame might look at first to just be the question of whether or not resentment is appropriate, but Strawson's framework allows this issue to be sharpened. The compatibilist and incompatibilist might disagree on whether the resentment is warranted, but they should be able to agree that it is fitting. What settles the question of whether or not resentment is appropriate is if the actions in question are truly manifestations of ill will or not. Since we have stipulated in this case that they are, resentment will be fitting whether or not the actions were causally inevitable. This is because, just like truth, fit depends on content, not on origins.

Understanding this fact about the fittingness conditions for reactive attitudes has put us in a position to see how and why there is a fallibility constraint in ethics. As we have seen, it cannot be one that bottoms out in the Principle of Alternate Possibilities. Instead, it comes from our confidence that the world truly contains ill will that is fitting to resent.

Think back to the epistemic case. Suppose we presented it to a philosopher who informed us that he had stopped listening once the case stipulated that the agent believed he was in a room with a cube, a cone, and a pyramid. When pressed to explain why, he replied, "Well, I knew that it was true as soon as you told me he believed it. A false belief is a metaphysical impossibility —it is part of what it *is* to be a belief to be true." Any of us who have ever had contact with false beliefs will be able to see that this theory of belief is deeply confused.

Suppose we presented this same philosopher with our practical case and discovered that his theory of intention was parallel to his theory of belief. Rejecting the case outright, he might complain, "If resentment is only fitting as a response to ill will, then I can be sure that it is never fitting, because to have a will is to have a good will. Supposing otherwise is a metaphysical impossibility. Part of what it is to have a will is to be benevolent." Once again, any of us who have experienced an instance of fitting resentment will immediately be able to see that such a theory of intentions and the will is also deeply confused. The connection between resentment and ill will is so deep it is likely much harder to take our imagined philosopher's verdict very seriously in this case. Even though we might often experience particular instances of resentment as unfit to the

circumstances in which we experience them, it is far rarer (although not impossible) to experience resentment itself as an unfit response to ill will.

Notice here that to take resentment to be fitting to ill will means that it is *not inappropriate* to feel it in response to ill will. It is not obligatory to feel, in fact, we might applaud the character of those who do not rise to resentment even when it is both warranted and fitting. But notice the way that resentment is a fitting response to ill will in a way that say, gratitude, is not.

As Strawson points out, for it to be an instance of resentment it will have to understand the agent it is resenting as bearing some ill will toward it, and so every instance of resentment carries with it a commitment to the actuality of ill will, which presupposes the possibility of ill will. Anyone who has evidence for a single instance of warranted resentment also has grounds for a fallibility constraint in ethics. If intentions could not fail to be kind, then no behavior could genuinely manifest ill intentions and no resentment could be either warranted or fit.

Depending on one's other philosophical commitments, this source of the fallibility constraint can provide even firmer ground than just that of an instance of fitting resentment. After all, whether or not one has experienced a genuine instance of fitting resentment is an empirical question, and demands access to facts (such as the relationship between other agent's behaviors and intentions) that we might never have. Pointing to actual instances of *warranted* resentment is really only necessary to address the doubts of those who are skeptical about the *fittingness* of resentment, which is to say its appropriateness in relation to ill will.[44]

[44] As I just mentioned, such skeptics are possible, but unlikely.

Think back to our discussion separating a sort of conceptual fallibility constraint from the Principle of Alternate Possibilities. There, we supposed that even in a world deterministically guaranteed to never realize a practically irrational action, the conceptual fallibility constraint would still be satisfied because there was still a stable conceptual category for practically irrational actions. That is to say there was still something that *counted* as irrational action, even if it never happened to be realized. Such a theory of practical rationality was able to satisfy a conceptual fallibility constraint because the conceptual particulars of the theory, not because of the metaphysical particulars of the world. Accepting that resentment fits ill will is enough to admit the *possibility* of a will that fails to be good, even if this possibility is never realized.

Strawson's discussion of resentment, ill will, and the roles they play in our moral universe allow us to see a precise analogue to the practical irrationality case. By drawing our attention to ill will, via our distinctively moral response to it, allows us to see that, whatever our answer to the empirical question of whether or not the world happens to contain any, ill will is a conceptually coherent entity, and any theory claiming otherwise will have made a mistake in doing so.

If one accepts the conceptual coherence of ill will, then this amounts to a proof of its conceptual possibility, and so should be enough to ground any complaint in moral philosophy (or elsewhere) that a theory has gone awry in denying this conceptual coherence.

However, arguments that appeal to bare conceptual definitions can do little to convince those who do not already accept the concepts in question, and there are at least a few philosophers (whom I will mention in the next subsection) who at least have the

incoherence of ill will as an apparent entailment of their other deliberate and substantive views in moral philosophy. These philosophers and their advocates might be inclined to push against the conceptual coherence of ill will, either to save other substantive doctrines of their own, or simply to cause problems for my present account.

To such skeptics about the conceptual coherence of ill will, Strawson's framework of reactive attitudes provides grounds for a particularly helpful reply. Strawson allows us to see that those who genuinely wish to deny that ill will is conceptually coherent and fits resentment will also be forced to deny other central features of the moral universe. Not only will they have to deny the fittingness relation between ill will and resentment, but they will have to deny the conceptual coherence of resentment itself. Since resentment conceptually involves ill will as its target, an incoherence in ill will will mean an incoherence in resentment itself. At this point, denial of any instance of resentment seems to come at even greater cost. There is a certain appeal to the idea that everyone means well whenever they act and there is genuinely no malice, only benevolence confused or misdirected.[45] This happy picture of human nature can appear a palatable consequence in some lights, but reflection on the Strawsonian reactive attitudes allows us to see that believing that everyone ultimately means well will undermine *any* instance we might have had to feel resentment toward anyone for anything they had done to us. And this seems quite a price. It is one thing to think that the world is basically a happy place. It is quite another to have to deny that one has ever been slighted or wronged.

[45] Elsewhere I wish to develop the case that Francis Hutcheson held a view very much like this.

Strawsonian reactive attitudes not only point out the conceptual coherence of ill will that any moral theory is at its peril to deny, but it also provides us with further resources to engage those who are skeptical of this very conceptual coherence.

What Strawson has allowed us to see is that there is a vindication of the fallibility constraint to be found in the realm of moral responsibility, but not in the most obvious place. In "Freedom and Resentment," Strawson claims that the truth of determinism is irrelevant to what is perhaps most important for moral responsibility, since it is the nature and direction of the intention, not its metaphysical determination, that makes it fitting to resent. The truth of determinism also turns out to be irrelevant for the justification of the conceptual fallibility constraint that is often at play in moral philosophy. We can see this because a conceptual commitment to infallibly good intentions can do what even the truth of determinism could not, and that is to threaten "the general framework of attitudes [which] itself is something we are given with the fact of human society." (FR 23) If ill will truly were conceptually impossible, our resentment would be an unjustified response to any behavior, however irresistibly we might continue to feel it.

Of course, our feeling of resentment and its fittingness in response to ill will is not perfect proof of either its actuality or its appropriateness relation, and this is not a problem my account can dismiss as irrelevant. As I indicated above, I hope to have drawn our attention to the conceptual coherence of ill will, and to those who persist in being skeptical of such coherence, I can further point out that our confidence in a moral fallibility constraint should be every bit as strong as our confidence that we have felt justified resentment or that resentment fits ill will.

My confidence that I have experienced at least one such instance is quite high, and should any reader's confidence be so low as to question my conclusion, I have some nasty things to say about his mother.

ii) What resentment gives us evidence for is a particular kind of culpable agent. What is distinctive about resenting someone is that one must simultaneously take that person to be both a full agent and to be behaving with intent to harm. Resentment carries with it the implication that the ill will comes from a full agent responsible for its attitudes. The resentment is a non-optional attitude if we are to see others as fellow agents. To think of the agent as less of a participant and as more of an "object" is to make it ineligible for resentment. While we might suffer from the hostility of creatures who are not full agents, who are less than participants with us, we cannot properly *resent* this malevolence without also thereby raising our estimation of that creature to be a full participant with us. This should push us to reject any moral theory that has it such that full agency cannot be malevolent. A theory that does so would dismiss as impossible an indispensible feature of our moral world.

It might seem deeply improbable that there would ever be a moral theory that does dismiss this possibility, and as a matter of explicit initial commitment, this is probably correct. This is why it usually takes an opponent of a view to establish that it is, in fact, committed to such a difficult thesis. But the threat of this commitment looms over the systematic views of several philosophers.

As I mentioned before, with Korsgaard's paradigmatic use of the fallibility constraint, Humean accounts of practical reason have been the target of just this sort of

criticism. Korsgaard's may be the most prominent use of the constraint, but it is not the only one. John Broome makes a cognate of the complaint in his "Can a Humean be Moderate?"[46]

Kantian theories of morality have also received similar criticisms. Versions of this worry have followed Kant's view across disparate traditions. Michelle Kosch has traced a strain of it in Schelling and Kierkegaard.[47] Henry Sidgwick was perhaps the first to make a version of the point in English in an 1888 paper.[48] In our own century, Douglas Lavin has raised this worry about Christine Korsgaard's Kantian view.

With the possible exception of Platonic understandings of practical agency, that insist that "to know the good is to do the good," thinking of goodness as a conceptually necessary (or even constitutive) part of a rational will is a theoretical commitment that no one sets out to have. It is worry that plagues at some of the most influential traditions in modern ethical theory, and, as I will show in the following chapters, it is a worry that will undermine at least one current attempt to ground normativity in contemporary moral philosophy (Christine Korsgaard's, see chapter 3), while revealing a hidden strength in a historical account not often taken seriously (Bishop Joseph Butler's, see chapter 4).

Finally, I must admit that the criticism in one of those pieces, Korsgaard's critique of the Humeans, is not directly justified by Strawsonian considerations about our commitment to fitting resentment. It is not strictly appropriate to *resent* people for their morally neutral irrationality, so being persuaded that there must be justified resentment will not also offer grounds to believe in irrationality.

---

[46] Broome, 1993.
[47] See Kosch, 2006.
[48] Sidgwick, 1888.

But this is actually a strength of the justification I have uncovered. Moral philosophy was probably the right place to begin the search for a justification of the fallibility constraint, especially because the Principle of Alternative Possibilities is such an appealing prospect for it, but the justification we have found need not be confined to ethics. It has a structural feature that allows it to be exported to other normative domains. Wherever we have compelling grounds to believe that there is or has been a negative instance of a kind, then we have found a justification for another, structurally identical fallibility constraint. Fully examining this is beyond the scope of my current project, but I will return to this matter with some orienting remarks at the end of Chapter 5.

iii) At one point in "The Normativity of Instrumental Reason," Korsgaard remarks that

> Hume's view seems to exclude the possibility that we could be *guided* by the instrumental principle. For how can you be guided by a principle when anything you do counts as following it? (NIR 229)

As we have seen in chapter 1, it is perfectly possible to be guided by a principle that you cannot violate. Instead of posing this rhetorical question, Korsgaard, along with the rest of us, must insist on the possibility of failure only when we have independent reason to believe in the actuality of failure. When we have such reason we will need norms that can acknowledge this failure and steer us away from it. This fact and this need is what vindicates the demand to leave open the possibility of failure, not considerations about the nature of rules or about the metaphysics of freedom.

Strawson's diagnosis of the mistake made by both the "pessimists" who think that moral responsibility requires disproving determinism and the "optimists" who disagree is

that "[b]oth seek, in different ways, to over-intellectualize the facts." (FR 23) The philosophers who have sought to use or ground the fallibility constraint, myself included, have also been prone to just this sort of over-intellectualizing. It need go no further, however, since all we need to anchor the fallibility constraint is to observe that actions are sometimes irrational and intentions are sometimes bad. And what justification could be more compelling than this?

Let me sum up what I take myself to have done in this chapter. I have suggested the plausibility of looking to moral responsibility as the place to find a vindication of the fallibility constraint in ethics. Furthermore, I have argued that the Principle of Alternative Possibilities is *not* the proper source for such an ethical fallibility constraint, however appealing it might have appeared. PAP can ground a kind of fallibility constraint if it is correct, but not the sort that can support the work built on what I have been calling the "fallibility constraint." Instead, I have claimed that our confidence that there is a fallibility constraint in ethics should only be as high as our confidence in the possibility of failed instances of a kind. In this case, it is wills that could fails to be good, and in this case, our confidence that there is such a thing should be quite high. With an ethical fallibility constraint so established, it is now time to see what further insights its application to actual moral theories can bring us.

Chapter 3

A Case Study in Constitutivism

I. Introduction

Now that we have reason to trust that the fallibility constraint should govern our thinking in ethics, we can apply it to various moral theories to see what it reveals. Constitutivism has been an influential movement in moral philosophy recently, and it is where I would like to start applying the fallibility constraint. To do this, I will examine a notable and recent version of the constitutivist view.

That is the version presented by Christine Korsgaard in some of her recent articles and lectures. Specifically, it is the view advanced in her recent series of Locke Lectures,[49] part of which has already appeared in her "Self-Constitution in the Ethics of Plato and Kant."[50] These considerations have since been revised and collected into her latest book, *Self-Constitution: Action, Identity, and Integrity*.[51] Let me flag at the outset that this chapter is based on her view as it was expressed in the one published paper and the electronic text of her Locke Lectures that she graciously made available on her website. I am nearly positive that my criticisms in this chapter will not hit their mark exactly because Korsgaard herself read an earlier version of this chapter, and restated part of her view in light of it. Unfortunately, the book was not published in time for me to

[49] Korsgaard, 2002. I will hereafter refer to them as LL N xx, where N is the Roman numeral of the number of the lecture, and xx is the page number of the .pdf version of the text.
[50] Korsgaard, 1999.
[51] Korsgaard, 2009.

revise my chapter in light of it, and so I only intend to move forward with this critique once I have a chance to compare it to the view to which she has committed herself in print. I am doubtful that the difficulties I point out here can be remedied through a simple restatement of her view, since they seem to be a product of its deepest philosophical machinery. But allow me to reiterate that this is merely a suspicion, brought in ignorance of how the view has changed since the Locke Lectures.[52]

As we saw in Chapter 2, Korsgaard is highly aware of the fallibility constraint and its applicability to moral philosophy; it is the basis for her argument against Humeans in *The Normativity of Instrumental Reason*. Furthermore, she recognizes the fallibility constraint as presenting a looming difficulty in the philosophical apparatus she inherits from Kant. She is one of the very few contemporary philosophers to have devoted a large amount of intellectual energy to resolving it in her own account.[53]

To present my analysis, I will argue in seven further sections. First, because I find the account so intuitively unappealing, I will offer a motivation for adopting it in my first section. This is the fact that views such as these promise an attractive position against (at least) two sorts of ubiquitous moral skepticism, all while satisfying an intuitively powerful requirement on all proposed moral theories. Next, because both the problem I am concerned with and her solution to it will be described in Korsgaard's own

---

[52] If it does turn out that Korsgaard's view has been altered, and specifically so as to better accommodate the fallibility constraint, I will be thrilled, since this is an object lesson in the power and importance of the fallibility constraint. If this is the case, I am content to aim my complaints here at the "earlier Korsgaard," even conceding that they were fixed in the "later Korsgaard." Though this runs counter to my general argument against constitutivism at the end of this chapter, it is not a possibility I should be so cavalier as to rule out before reading the new book.

[53] In Lavin, 2004, Douglas Lavin has already made a version of the point I make below, that Korsgaard's commitments cause an incoherence in her overall view. This paper was published after I had written much of this chapter, and I hope that my discussion of the analogy to Plato and my earlier analysis of the fallibility constraint are enough to make it an independently interesting addition to Lavin's very helpful work.

terms, I will have to invert my presentation somewhat, by presenting enough of Korsgaard's view in my second section to generate the problem, and then describing the problem it is meant to be a solution to in my third. My fourth section will describe the rest of Korsgaard's view that is designed to resolve this problem. In section five, I will present my argument that Korsgaard's view does indeed offer a solution to the stated problem. But this solution is a solution in name alone, and does not come in the way Korsgaard supposes it does, nor in a way that can maintain its attractive position against either sort of moral skeptic. Then, I will discuss how the way that Korsgaard manages to solve the "mereological problem" she sets for herself prevents her from grounding moral demands from this solution as she intends. Finally, I will conclude arguments to generalize the problem that faces Korsgaard's view to all constitutivist ethical views generally.

II. Two Forms of Moral Skepticism and the Categoricity Requirement

Korsgaard offers a "constitutive account" of practical and moral action. This means, roughly, that she presents the standards that must be met for something to count as an action (i.e. the criteria that *constitute* action) and then derives practical necessity and moral normativity from these standards. It is not my insight that such a strategy supplies answers to skeptics. Korsgaard herself notes this virtue in her first Locke Lecture: "The idea of a constitutive standard is an important one, for constitutive standards meet skeptical challenges to their authority with ease." (LL I 21) To skeptics who ask "Why must I ξ in order to φ?" the reply is simple and direct: "Because ξing is what it is to be φing. Sure, you might fail to ξ, but then you would have to be doing

something other than φing."[54] Imagine a skeptic who asks why she must put pigment on the walls in order to paint them. The answer is simple: because failing to do so would be failing to paint the walls.

But this sort of immunity to skepticism is purely hypothetical. *If* you are to φ, *then* you must ξ. But when the φing in question is moral action, the skeptic is not concerned with what it takes to be moral, the skeptic is instead asking why one should φ in the first place. Even though she does not make this entirely explicit, Korsgaard's constitutive account still promises an answer to at least two kinds of skepticism and one intuitive requirement on morality. I will call these two versions "naïve" and "sophisticated", in reference to how much background in moral philosophy they presuppose. I will call the intuitive requirement the "categoricity" requirement, since it is the idea that moral demands are categorical in applying equally to all agents in relevantly similar circumstances. I will first describe them and then point out how Korsgaard's constitutive account promises to meet them.

What I will call "naïve" skepticism is the most ground-level sort of skepticism about morality. It squints doubtfully at any proposed moral claim or potential moral value and simply says "Say I don't like that claim and I won't respect it. Why should I?" or "Say I don't value that and I don't want to come to. Why must I?" This sort of skepticism recognizes the queer normativity morality claims to have over even those who

[54] Here, I'm deliberately paraphrasing David Velleman when he presents his own constitutive account of action in "The Possibility of Practical Reason" pp. 170-199 in the book of the same name. See Velleman, 2000.

do not actively endorse it, and, quite legitimately, questions its authority to do so. It is simply the skepticism that asks, "Why be moral?"[55]

The more sophisticated form of moral skepticism is such because it comes from a substantive view in moral psychology.[56] Though the historical accuracy of the title can be disputed, the moral psychology is usually called "Humean." This view furnishes the mind with two kinds of entities, beliefs and desires. Beliefs concern matters of fact and relations of ideas, are answerable to the world for their correctness, provide no motivational force, and are subject to the authority of reason. Reasoning may revise existing beliefs or discover new things to believe. Desires may also concern matters of fact, but they are not answerable to the world for their correctness, since they prescribe how the world is to be. Perhaps because of this, they are not subject to reason's demands. Instead, they command it to supply them with the information about how to make the world fit their descriptions. And as such, they are the only source of motivations for action.

A more sophisticated sort of moral skepticism can arise against this backdrop of moral psychology. If moral values, of any sort, were to exist in the world in any way, we could only come to have beliefs about them, since only beliefs are answerable to the world for their correctness. But beliefs provide no motivational force themselves, and need not impact the desires that can provide such force. Desires are not subject to reason or the world. Thus, even if we could come to know what was moral and immoral, we

---

[55] Pitching these constitutive accounts as replies to skeptics is at least as old as the *Republic*, but I must confess I would likely not have thought to consider constitutive accounts in light of this skepticism if not for the character Gary in Peter Railton's "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action", See Railton, 1997, pp. 53-79.

[56] I owe much of what understanding I have of this version of moral skepticism to Michael Smith's Introduction to his *The Moral Problem* (Smith, 1994).

would not necessarily be moved to act in accord with them, because our source of motivation is wholly severed from our moral sensitivity. Surely, we might coincidentally desire some or all of that which is moral and applaud this happy accident, but this would be pure serendipity —it would be a grave mistake to suppose our desires somehow became sensitive to our moral beliefs. In a way, this skepticism is even more powerful than the general naïve skepticism. It can allow its advocates to be convinced that some course of action is moral or good, even endorse the most robust sort of moral realism, and still contend that there is still no reason to be moral, since no number of true beliefs about the world can coalesce into a desire and thus a reason to act. Finally, Korsgaard's constitutive account promises to meet the categoricity requirement on moral thinking because it will apply to all agents in virtue of their very agency itself.

If successful, Korsgaard has compelling responses to both of these common forms of moral skepticism and a way of automatically satisfying the categoricity requirement. Her strategy is to work out what is necessary for agency, what its constitutive standards are, and then show that moral actions are deducible from the simple requirements of what one must do in order to be an agent. So far, this grounds only a hypothetical normativity (of the form: If you are to φ, then you must ξ.) which can be rejected.

To solve this problem she will appeal to the necessity of action, the fact that one must still deliberate freely about what to do, whatever the metaphysical realities are. Korsgaard can claim that one must think of oneself as a free agent, and thus must decide to act. If the skeptic agrees, there is no further argument needed, and her project can proceed. If the skeptic disagrees, he can only do so because he is thinking of himself as free to disagree, and her project can proceed. Whatever the skeptics claim, they must be

agents, because they must act.[57] Her argument proceeds as a constructive dilemma, deducing that moral normativity has a grip on all agents, categorically, i.e. in virtue of their agency itself, and so agents are committed to morality. An agent may opt to agree that she is free or not; but, in either case, so opting entails that the agent in fact takes herself to be free. Agents, in virtue of being agents, are required to choose. If moral requirements do indeed follow from this constitutive fact about agents, then no agent (skeptic or no skeptic) will be able to legitimately avoid the demands of morality.

If these claims could be established, it seems that both forms of skepticism could be dispelled. Sophisticated skepticism would be shown to have made a factual error in claiming that only desires can have motivational force. The necessity of action commands that we do *something*, whatever our desires are, even if they are nothing at all. Further, it claims that once we reason about what it is to act and notice that we must act, we will have reason to perform some actions in particular. If this is correct, then reason and belief can recommend particular actions, as well as supplying the motivational force to drive them (derived from our predicament in the necessity of action). The sophisticated skeptic will have to concede, since the picture of moral psychology his skepticism depended on will be shown to be mistaken.

Naïve skeptics would also have to concede that they have reason to be moral and cannot even cry foul. Any attempt to obligate them to be moral or to appeal to the value of being moral in order to resolve these skeptics' questions must immediately fail. It is simply begging the question to appeal to morality in an effort to establish it. But Korsgaard's strategy relies on the morally neutral concept of agency. Unlike Gary, the

[57] We might well doubt that action really is inescapable for humans. (David Velleman worries as much in Velleman, 2006.) But since the truth of this premise is tangential to my interests here, I will only register this worry and not explore it further.

skeptic in Railton's "On the Hypothetical and non-Hypothetical"[58] the naïve skeptic does not doubt action, just moral action.[59] If Korsgaard can show that moral normativity falls directly out of the necessity and nature of agency itself, she will have convinced this skeptic as well.

Thus, despite the fact that people like me find views such as Korsgaard's counterintuitive both initially and in their implications, there is much to be gained against two common and compelling forms of moral skepticism. Korsgaard's account seems worth the effort to defend, so it is now time to see how she defends it.

III. Owning Actions, Particularistic Willing, and Mereology

Korsgaard's argument can be thought of as a constructive dilemma, with the necessity of action as its first disjunctive premise. If it can be grounded, the necessity of action will spill over into any of its essential attributes, proving them to be necessary as well. It is the observation she begins her Locke Lectures with, so it is a sensible place to begin my sketch of her account. As she puts it

> Human beings are *condemned* to choice and action. Maybe you think you can avoid it, by resolutely standing still, refusing to act, refusing to move. But it's no use, for that will be something you have chosen to do, and then you will have acted after all. Choosing not to act makes not acting a kind of action, makes it something that you do. (LL I 1)

Whatever the metaphysical status of our freedom, we must act under the idea of freedom and choose some course of action or other, and think of ourselves as the cause of that action. Whether we concede this point or attempt to deny it, the only way we can possibly pursue either course of action is by considering ourselves to be free to do so.

---

[58] Railton, 1997.

[59] Korsgaard's account promises to convince even skeptics like Railton's Gary, who, in order to be skeptics, must be agents.

Whatever else the skeptics may claim, they cannot reasonably deny this "unconditional" necessity to act whose necessity is "not causal, logical, or rational necessity. It is our plight: the simple inexorable fact of the human condition." (LL I 2)

In addition to offering its potential answers to the skeptics, Korsgaard's view offers a way to distinguish actions from events. She does this in what she calls "the argument against particularistic willing" which is meant to show what is necessary for an action to be willed by an agent (these features themselves will then be necessary, courtesy of the necessity of action). The argument is fundamental to Korsgaard's entire view, since she derives the rest of her conclusions from it.

I will present the argument's bare skeletal premises first, and then explain how they fit together once they have been put on the table. As far as I understand it, the argument runs as follows:[60]

> [1] All willing [of actions] is either universal or particularistic. Since there is necessarily action, if it cannot be one, it must be the other. (premise)
>
> [2] To conceive of yourself as the cause of your actions, you must identify with something in the scenario that gives rise to the action, and that is your principle of choice. (premise)[61]
>
> [3] To will particularistically is to will on the basis of no principle at all; it is a will indistinguishable from the collections of inclinations at play in you. (premise/definition)
>
> [4] Having a particularistic will is just having no will at all. Willing requires principles with which one can identify, and

[60] The following is a paraphrase and interpretation of two nearly verbatim passages that appear in SC 26-27 and LL II 19-25.

[61] As I will note later, this is crucial if what I describe as the mereological problem is to be solved, but it is still obscure to me why Korsgaard here thinks that only identification with a *principle* is capable of solving the mereological problem. Another way of putting the criticism I detail below is to notice that mere identification with *anything* is enough to solve the mereological problem, and so solving it doesn't recommend any particular course of action, let alone only the moral ones.

particularistic wills are those that are, *ex hypothesi*, those without principles. (from [2] and [3])

[C] Particularistic willing is impossible. All willing must be universal. (from [1] and [4])

At the very least, the argument looks valid. We now need to know what its terms mean so that we might judge its soundness. What is most important is to understand is what it means to will either universally or particularistically.

If particularistic willing were possible, Korsgaard says, "it would be possible to have a reason which applies only to the case before you, and has no implications for any other case." (SC 23-24) To will particularistically, you will have to identify fully with the inclination that prompts your action. Further, you would have to see that inclination as wholly unique and metaphysically particular. For, "if you had a particularistic will you would not identify with the incentive as representative of any sort of type, since if you took it as representative of a type you would be taking it as universal." (SC 26) Such willing excludes the possibility of appealing to a principle, because principles speak in the language of universals. Principles dictate particular types of actions in particular types of circumstances. Even if they are so heavily specified as to only be invoked once, they still prescribe that one perform a *type* of action when one encounters a *token* of a type of circumstance. Even if there turns out to be only one such token circumstance, the principle that determines action in that circumstance implies a precedent. Thus particularistic willing for particularistic reasons can be thought to set no precedents, appeal to no principles, and speak in no universal terms.

Premise [1] tells us that particularistic and universal willing exhaust all the sorts of willing there is. The easiest way for this to occur is if all willing is thought to be either

particularistic or not (p v ~p) —this will be true analytically. And universal willing is just what we would expect it to be, that is to say not what particularistic willing is. Universal willing is the willing of an action that is thought to have implications for future cases, to set precedents for circumstances of the same type, and to freely use universal terms. It is especially important to notice that this sort of universal willing, the sort that is the antithesis to particularistic willing, is not restricted to a Kantian framework. Korsgaard is quite explicit about this, noting that "Someone who takes 'I shall do the things I am inclined to do, whatever they might be' as his maxim has adopted a universal principle, not a particular one: he has the principle of treating his inclinations *as such* as reasons." (SC 27)

Premise [2] can be understood as follows: If action does indeed require that you conceive of yourself as causally responsible for the events that occur, Korsgaard is right to claim that you must identify with something in the action, and that thing had better be responsible for the action. To illustrate this, consider Korsgaard's own example of an action with only three elements: a desire to A, a desire to B, and a principle preferring A to B.[62] Imagine that you A on the basis of the principle –in that case it was the principle that was responsible for the action. If you identify with this principle, you can claim the action as your own because you are thereby also responsible for the action.[63] Should you fail to identify with the principle that was responsible for the action, then it is as if you are a mere spectator to your own actions. If you were not responsible for the action, then

[62] See SC 26.

[63] I am not at all clear on how deep the metaphysics of this "identification" are supposed to run. For the argument to work, it seems that you must consider yourself to be strictly identical with the principle, so that you do not suspect you are some fourth, passive element in the action. But it seems clear that Korsgaard is relying on a more colloquial sense of "identify", as in the way one might identify with a political cause or dramatic character.

it could have gone on without you, and it was hardly yours to be authored. If your actions can come as the result of principles, and those actions are to be yours, you must identify with those principles.

We are now in a position to understand why particularistic willing is impossible. It is supposedly a kind of willing that cannot possibly involve principles. Therefore, there can be no principles to identify with in the action. At best, one will have to identify with the causally efficient desire or inclination responsible for the action. And, for Korsgaard, this is the same as having no will at all:

> Particularistic willing eradicates the distinction between a person and the incentives on which he acts. But then there is nothing left here that is the *person*, the agent, that is his will as distinct from the play of incentives within him. He is not one person, but a series, a mere conglomeration, of unrelated impulses. There is no difference between someone who has a particularistic will and someone who has no will at all. Particularistic willing lacks a subject, a person who is the cause of these actions. So particularistic willing isn't willing at all. (SC 27) [64]

The way Korsgaard proposes we distinguish actions from events is intuitive enough: actions are those events that a person is causally responsible for. Events lacking persons to cause them are merely events, not authored by agency. More specifically, actions are those events performed on the basis of universal reasons, because particularistic reasons (could such things exist) would simply lack agents to perform them.

Understanding this way in which actions are distinguished from events sheds some light on Korsgaard's frequent but sometimes cryptic catchphrase that "action is self-constitution." Because when you truly act, you present a principle with which you

[64] I am led to wonder why identification with a principle leaves something else for an agent to be in an action, when identification with a desire does not. Though Korsgaard finds identification with a desire to be equal to having no will at all, it seems that other prominent theorists, like Harry Frankfurt, take this to be precisely what a will is. But, again, that is a different paper.

identify and you thereby offer your identity into the world. If you fail to do this it seems that there is nothing to identify you with (or identify as you), and you are nowhere to be found. It is only in acting that you have an identity, in a far more than metaphorical sense. It is because she thinks action points out where and what you are in the world that she says things like "So whatever else you are doing when you choose a deliberative action, you are also unifying yourself into a person." (SC 27)

Before moving on, we should pause to notice a concern that Korsgaard continually returns to throughout the Locke Lectures, and is a clear motivation for the rest of her account. It functions as a premise and explanation at several points in Korsgaard's reasoning, and I will call it "the mereological problem." She registers it in the very first Locke Lecture by saying

> I also believe it is essential to the concept of agency that an agent be unified. That is to say: to regard some movement of my mind or my body as *my action*, I must see it as an expression of my self as a whole, rather than as a product of some force that is at work *on* me or *in* me. Movements that result from forces working *on* me or *in* me constitute things that happen to me… (LL I 15)

One puzzle about agency that Korsgaard is interested in solving is fundamentally a mereological problem. She wonders how it is that an agent could coalesce up out of the "in Aristotle's wonderful phrase, *mere heaps*" of psychological urges and materials. (LL I 19) For Korsgaard, this is a perfectly general and decidedly mereological problem.

Heaps of beliefs and desires become agents in precisely the same way that heaps of bricks and plaster become houses: through teleological organization. (LL I 20)

In order to find an agent in (or, as Korsgaard would put it, "over and above") the "mere heap" of psychological building materials, Korsgaard believes that we must find a

principle that unifies and organizes that heap into an agent, only then can we find genuine actions and genuine agency.

IV. No Bad Actions, No Bad Agents

As I argued in Chapter 2, a moral theory must leave open the possibility of failure if it is to be able to accommodate a crucial piece of data from our moral life. As I mentioned in the Foreword, Kant was thought to encounter this problem in a very specific way, and it is now time to briefly revisit his version of the problem and how Korsgaard's account inherits and exacerbates this problem.

In Kant's moral philosophy, the thing that distinguished actions from events and that gave them moral worth was the very same feature: autonomy. When an "action" failed to be autonomous, the resulting event was explained by an external force (or desire, inclination, etc.) that was forcing the agent's hand. Given this picture it looks like there is no way to properly resent the ill will in someone's bad actions directed toward you, since there are no bad actions. Because the presence of autonomy both makes an event an action and confers on that action moral value, these two features of an event will covary. Thus, when an agent truly acts (i.e. exercises her autonomy to be causally responsible for the events that occur), these actions will necessarily have moral worth. Actions that lack moral worth cannot properly be called "actions." Their lack of moral worth derives from the fact they were performed heteronomously (i.e. caused by some force alien to the agent's will). But heteronomous "actions" are mere events, caused by something other than the agent. Thus, on this view, there appear to be no such things as bad actions; only good actions or bad events.

Because we have done no work to ontologically eliminate her, it is natural enough to assume that the agent is still present in her heteronomous "actions," at least on Kant's view of things.[65] As you likely will have guessed from the argument against particularistic willing above, what Korsgaard adds to this picture is the ontological elimination. For her, at least in the context of action, persons are nothing more than the efficient principles of choice with which they identify.[66] Heteronomous "actions" include no *efficient* principles of action, since they are just cases of inclinations usurping control of the action. Since there is nothing for the agent to identify with in heteronomous action, she does not figure in to the explanation of the action. There is nowhere for the agent to be present in "actions" such as these, so Korsgaard concludes that the agent is not in them. In this way, Korsgaard inherits Kant's problem and does him one better by apparently committing herself to the claim that not only are bad actions impossible, but so are bad agents. Heteronomous "actions" are not universal, so if they are anything, they are particularistic, but you cannot will particularistically or be a particularistic agent, because doing so destroys agency by reducing it a mere heap of psychological urges. As Korsgaard says, "A truly particularistic will must embrace the incentive in its full particularity: it, in no way that is further describable, is the law of such a will." (LL II 24) But this will mean, as Korsgaard says

> that particularistic willing eradicates the distinction between a person and the incentives on which he acts. But then there is nothing left here that is the *person*, the agent, that is his self-determined will as distinct from the play of incentives within him. He is not one person, but a series, a *mere heap*, of unrelated impulses. (LL II 24)

[65] Although, as we will see in Chapter 4, Sidgwick's understanding of the problem for Kant has much higher stakes.

[66] For Korsgaard, they must be such if the mereological problem is to be solved, and there is to be an agent instead of a mere heap of psychological urges.

V. The Possibility of Bad Actions and Moral Normativity

As I said before, much of what attracted me to Korsgaard's discussion of this problem was that she recognizes this problem as a pressing one and then offers a solution to it. She recognizes that the constitutive accounts such as Plato's, Kant's, and her own claim that the metaphysical property that makes actions what they are also implies a normative property. I am happy to consider this a reason to discard these accounts, but Korsgaard entirely disagrees: "Now rather than finding in this a reason for rejecting these arguments, I think we should see it as our main reason for embracing them." (SC 14) I take it that this is because these views make all the promises I have noted above, and because Korsgaard believes that these views do in fact allow for bad action and can ground moral normativity. It remains to be seen how this can be done.[67]

i) Bad Actions

If action really just is self-constitution, then one might well make the case that the bad actions are those that do a bad job of constituting the self. This is exactly what Korsgaard does.

As has been claimed in the argument against particularistic willing, in order for even the possibility of an agent to be present in an event, there must be a principle of choice with which the agent can identify herself. But different principles will have different amounts of success in constituting a unified self. In every case that we identify ourselves with some principle of choice or other, we can be said to be truly acting, in Korsgaard's sense. What enables us to judge some of these actions to be good or bad is

[67] As we saw in Chapter 2, I disagree with Korsgaard as to *why* her account is under pressure to leave bad action possible, but we both will find her account to be subject to some kind of fallibility constraint.

the fact that we can evaluate how well they succeed in their constitutive aim of self-constitution and rank them accordingly.

It will be easier to understand why if we take a moment to examine one of Korsgaard's inspirations for this move, the rankings of different forms of government in books XIII and IX of the *Republic*. There, Plato ranks the five forms of government (and soul) from best to worst: aristocracy, timocracy, oligarchy, democracy and tyranny.[68] They are all recognizable as forms of government, but we can still evaluate them in terms of how well they meet their constitutive aim of governance. Korsgaard examines each one and points out that, the farther we move down the scale of bad governments, the more likely the governments are to break down. Aristocracies are thought to be the most unified, able to be flexible and maintain order even in times of crisis. Timocracies generally fare well, since valuing honor and victory will take a state far, but when these values conflict with the good of the state itself, the government risks breakdown. Oligarchies operate on principles of caution and prudence aiming toward some sort of mild hedonism. Korsgaard claims that such a plan is even worse for maintaining a unified state because of both the difficulty of determining what the satisfaction of the state's desires will consist in and the difficulty in trying to make them coherent. Democracy Korsgaard takes to be analogous to wantonness, because it makes no effort to make its desires cohere, and the coherence of its actions and constitution will be "completely dependent on the accidental coherence of [its] desires." (LL V 22) Tyrannies rank the lowest, but are strangely unified. The reason they are the worst sorts of government for states or souls is because their unification comes from a single, all-consuming desire that they organize everything around and for which they are willing to

[68] See Plato, 1997.

sacrifice everything, including their own stability. Tyrannies are simply what happens in a democracy if there is a single, overwhelmingly powerful force. They are so rigidly organized around this singular desire that they become brittle, and are thus the worst sort of constitution, because the unification they promise is fragile and fleeting.[69]

Korsgaard thinks principles of choice can be ranked in just the same way. The robust formulation of the categorical imperative is the best principle of choice in terms of self-constitution, presumably because it tells you to will only those maxims that can be universally accepted (by your future selves and fellow rational agents). Thus, you will only commit yourself to those things that can have universal support and are thus least likely to cause disharmony in the soul or civil unrest in the government. The actions that come about from an agent who identifies with the categorical imperative are thus the best sorts of actions, because they best achieve the constitutive aim of action: constituting the self into a unified whole.

But agents might identify with other principles of choice that constitute the self only partially, and this is how bad action is possible. Such actions are identifiable as actions, because they are made on the basis of principles with which the agent has identified, but they are defective in some way or another because they fail to wholly constitute the self. In Lecture Five, Korsgaard invites us to consider a bad agent performing bad actions as an agent who identifies with the principle of self love. Korsgaard mentioned this principle before (in the argument against particularistic willing) as one that might appear to be particularistic but is in fact universal, and it

[69] It is an interesting question that I will not presently explore whether the Categorical Imperative will, in fact, generate the most stability in an agent. Frankfurt is a promising foil for this conclusion, since he both worries that some commitments are so strong as to make an agent brittle (as he finds Agamemnon to be when he must decide whether or not to sacrifice Iphigenia) but also insists on the importance of the stable commitments he describes as "unthinkable actions," as we saw in Chapter 1.

merely instructs the agent to choose as inclination prompts. Though such a principle is universal and autonomous, Korsgaard believes it is so in only the weakest and most technical sense. A will that always constitutes itself around the appeal to another entity is autonomous, but in a defective way. Such a will is like Harriet in Jane Austen's *Emma*, whose principle of choice is to do whatever Emma thinks Harriet should do. There is a principle at work here, but this principle wins the title "autonomous" on a mere technicality: it is not really what we think of as an autonomous agent choosing for herself.[70]

Thus, it is in the adopting of defective principles that bad action is possible. The better the principle in producing actions that unify the self, the better the agents who adopt it and the actions they perform. These considerations lead Korsgaard to another catchphrase –that choosing bad actions is not a different activity from choosing good ones; it is the same activity, badly done.[71]

The necessity of solving the mereological problem is crucial in establishing the necessity of universal willing and thus establishing a normative force that requires the willing maxims as universal laws that extends universally, to all agents. In this way, Korsgaard's account can be seen as aiming to meet the categoricity requirement, by pointing at a feature that all agents will share, and necessarily so.

But solving the mereological problem is also what enables Korsgaard to accommodate the idea of *bad* actions. Bad actions, for Korsgaard, are simply those

[70] This paragraph paraphrases LL V 15-17 and SC 15-17. It is also unclear if this action, which is genuinely universal by Korsgaard's lights is capable of solving the mereological problem. An agent who wills to do just what his heap of inclinations would have done anyway is *conceptually* distinguishable from that mere heap, but perhaps only conceptually so. It is not clear that this is enough to solve the mereological problem.
[71] See SC 15 and LL I 21.

actions that are poor solutions to the mereological problem. In Lecture 5, following Plato, Korsgaard explains that "The kind of practical deliberation that issues in bad action is not a different activity from the kind of practical deliberation that issues in good action. It is the same activity, badly done." (LL IV 28)

Since all actions, in order to count as actions are solutions to the mereological problem and so unify an agent out of the mere heap of psychological materials that comprise him, Korsgaard thinks it is perfectly natural that we can therefore judge and rank universal principles of action in terms of how well, how harmoniously, and how sturdily those principles constitute unified agents. It is easy enough to see why a principle that brings an agent up out of the psychological constituents and is able to keep him there is better than one whose unification is merely fleeting or unstable. If all actions just are solutions to the mereological problem, it is clear why we would think that the solutions that last longer are therefore better solutions than those that are brittle and fleeting. Since these solutions are just actions, it is also intuitively appealing to see that we can judge the actions that constitute more stable and unified agents to thereby be *better actions* than those that do solve the mereological problem, but less lastingly. Provided that the Categorical Imperative generates actions and agents of a more stable sort than any other principle that generates actions, it will give Korsgaard a way to also accommodate the fallibility constraint. If other actions that come from other principles do what actions from the Categorical Imperative do, only less well, this provides a coherent description of actions which are fully actions (they unify agents according to principles and do solve the mereological problem) but are still degenerate instances of their kind (while they do unify agents, they do so only fleetingly or temporarily).

Therefore agents can perform good actions (or, perhaps more properly *best* actions) that come from the Categorical Imperative, but they are still capable of performing bad (or *worse)* actions that come from other, less successful principles.

It is in this sense that Korsgaard means that "action is self-constitution," because acting is what solves the mereological problem and constitutes an agent where before there was a mere heap.

ii) Moral Normativity

My current discussion concerns primarily the possibility of bad action, which Korsgaard argues for in the way outlined above. However, since her account is motivated in terms of grounding a moral theory out of the sheer concept of agency, it is worth a moment to see how Korsgaard thinks this follows. The sketch below will be brief and rough, but should be enough to see how Korsgaard intends her account to conclude.

Moral considerations are grounded due to the fact that "[t]he requirements for unifying your agency internally are the same as the requirements for unifying your agency with that of others." (LL VI 23-24) We have already seen that we must act, and that in order to act we must constitute ourselves. If what is required in constituting ourselves also requires us to unify our agency with others, and this is what morality amounts to, then moral normativity will be grounded.

Remember that the reason the Categorical Imperative was the best principle of self-constitution was because it advocated that you adopt only those maxims that you can will universally, and that will have maximum endorsement from others (notably your future selves) and thus will best unify the self. To this effect, Korsgaard says

> Constituting your own agency is a matter of choosing only 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, one you can live with later. (LL VI 24)

What this means is that when you choose your maxims, you must take the interests, concerns, and, in short, the "humanity" of your future selves into account, if you have chosen the Categorical Imperative as your principle of choice. Agency of this sort requires concern for others (your future selves), "[a]nd then one day it occurs to you that there is really no reason for you to treat yourself [or your future selves] differently than anybody else." (LL VI 5)[72] Therefore, you must choose actions that can be universally endorsed by all of humanity, not just all of your future selves.

Summing all these considerations we find that 1) we must act, 2) to act we must constitute ourselves, and 3) what is required in self-constitution entails unifying our wills with others. Therefore, from the mere fact of our agency and its essential structure, we find that we have reason to respect all the sorts of moral considerations that come from reasoning in light of the ends of others: benevolence, the cultivation of talents, a prohibition against lying and all the other usual Kantian conclusions.

I am deliberately neglecting to fill in the details of precisely how this is meant to follow, but these are the major stops on the final leg of Korsgaard's journey from the necessity of agency to moral normativity.

VI. Substantive Objections

i) Substantive vs. Procedural Justice

[72]This quote is taken out of context, but I believe it does no violence to the broad outline of her argument here.

If Korsgaard is correct, she has grounded moral normativity out of the very concept of agency and has overcome both of the skeptical challenges described in section II. But it is especially important to notice what sort of normativity it is that may have been grounded here. To see what sort it is, we need to take a moment to understand both what Korsgaard's distinction between substantive and procedural justice is, and what her commitments are in light of it.

When Korsgaard says "…usurping the office of another in the constitutional procedures for collective action is *precisely* what we mean by injustice, or at least it is one thing we mean…" (SC 7) she is employing the notion of procedural justice. This notion of justice declares an action to be just if and only if it comes as result of the proper following of the given constitutional procedures. Thus, in the United States a given law is procedurally just if it is passed by Congress, endorsed by the President, and deemed constitutional by the Supreme Court, because these are the procedures by which the United States makes its laws. An upset in these procedures will make the given law procedurally unjust. If Congress tries to enforce a law that the President has vetoed and the Supreme Court has struck down, then it is usurping the office of another and is performing an act of procedural injustice and thereby making the given law procedurally unjust.

Substantive justice, on the other hand, concerns our "independent idea of what outcome the procedures ought to generate" (LL V 9) and is the other thing we mean by "justice." Notice that according the United States procedures, *almost anything at all* can be a procedurally just law. If it has the proper support from the three branches of government coming in the proper way, it is a procedurally just law, almost regardless of

its content.[73] If you find this consequence unnerving or unjust in the least, then you must be appealing to a notion of substantive justice, because such a law is procedurally unimpeachable. A notion of substantive justice is what allows us to declare, say, a law banning half the population from voting on the basis of gender unjust, even if it has been procedurally endorsed.

Notice that these two sorts of justice generate two sorts of normativity that can easily come apart and push us in different directions. Procedural normativity might well urge us to φ, because our procedures have dictated it. But, we find φing to be substantively wrong or unjust, and thus our sense of substantive normativity urges us not to φ.

Once we recognize this tension, we might be tempted to tailor our procedures so as to always (or at least usually) dictate what is substantively just. But Korsgaard insists that we cannot do this, that procedural normativity derives its normative force independent of our notions of substantive conclusions.

> [I]f the normativity of our procedures came from the substantive quality of their outcomes, we'd be prepared to set those procedures aside when we knew that their outcomes were going to be poor ones. And as I've been saying, we don't do that. Where constitutional procedures are in place, substantive rightness, goodness, bestness, or even justice is neither necessary nor sufficient for the normative standing of their outcomes. (SC 9)

According to Korsgaard, procedural normativity binds us independently of substantive constraints, and it is a mistake to think otherwise.

---

[73] Of course, we can invent paradoxical content that cannot succeed in being procedurally just. All we need to do is have Congress try to pass a law that is stipulated to only go into effect once it receives a presidential veto, but this sort of indexically paradoxical law is not the sort we should be troubled by since it could hardly be the source of the kind of moral requirements we seek from a moral theory. If the content of morality is only what cannot be procedurally just in this way, morality will be radically, unrecognizably, different than we will have supposed.

As far as I understand it and have described it above, the moral normativity that Korsgaard aims at grounding in the Locke Lectures and elsewhere is a purely procedural sort of normativity. It binds us because we must act, and in order to act we must follow certain procedures. Fortunately enough, what is required by these procedures will overlap with much of what we take to be substantively good and just. This overlap, however, is not why we must abide by these procedures. They would have their grip on us regardless of what they required, because their necessity comes from the necessity of action, not the normativity of substantive considerations.[74]

Though she is deeply committed to the distinction between substantive and procedural considerations and the separateness of their respective sorts of normativity, she sometimes waffles on this point.[75] In response to the objection that we would surely abandon our governmental procedures if they generated substantively bad outcomes too often, she says "perhaps we should say that the normativity of the procedures comes from the usual quality of their outcomes *combined* with the fact that we must have some such procedures, and we must stand by their results." (SC 9) This is certainly a good description of what happens in Plato's account, and in a constitutional democracy such as the United States, but it *cannot* be a description of what happens in Korsgaard's account, or else she will violate her own constraints and lose her position against the naïve skeptic. It will be easier to say why if we understand why Plato is able to advocate such a mixed account.

---

[74] I should also note here that this sort of claim is precisely what Gideon Rosen claims cannot be done by constitutive standards. See Rosen, Manuscript, and Chapter 5 below.

[75] This waffling occurs in the text provided in the electronic version of the Locke Lectures. In person, Korsgaard clarified the point to me and has presumably addressed this issue in the book version of these lectures.

ii) The Platonic Hierarchy

Quite simply, Plato can evaluate the different forms of government in both procedural and substantive terms because he has perhaps the most robust metaphysics of the good in all of Western philosophy. What is substantively good is just what most participates in the ideal form of The Good, and it is no coincidence Plato's hierarchy of governments tracks not only their stability, but their ability to track this good as well. It even seems plausible to think that their various amounts of stability are due to their various abilities to track The Good.

Aristocracies are quite literally governments of the best, that is the best positioned to know what the good is. These governments are run by philosopher-kings, who are the only citizens in a position to have access to the form of The Good. Presumably, this is why they are always able to make the right decision for the good of the city, because they have knowledge of what that good is.

Timocracies and Oligarchies track other values that often coincide with what the good of the city is, but they will give the wrong answer about what is best for the city when Honor or Prudence do not coincide with The Good. Democracies do not consistently track any value at all; they merely reflect the will of the masses, which is fluid and inconsistent enough to sometimes accidentally coincide with what is good for the city. The only way to do worse than this is under a Tyranny, which only tracks the desires of its single leader, which very rarely (if ever) coincides with the good of the city.

Indeed, Korsgaard is correct to say that constitutional procedures are required for any collective actions of the state at all, but there are many ways to perform collective actions. A state's constitution might dictate that the auxiliaries willingly follow the

dictates of the philosopher-kings. Equally, a state could have a constitution that decrees that the state will do as the tyrant says, or else. The way Plato is able to rank these equally constitutional constitutions is in terms of how well they track what is substantively good. And this move is open to Plato because he already has a substantive metaphysics of the good to which he may appeal.

iii) Korsgaard's Hierarchy and Underdetermined Practical Necessity

Though at one point she says that perhaps the best way to choose or rank our principles will include a consideration of their substantive outcomes, this is not in fact the way she intends to rank the principles of choice she considers, and this is to her credit, for otherwise the naïve skeptic can cry foul. The very strength of her approach against what I called "naïve skepticism" in the first section is that it does not depend on any substantive values to ground its normativity. These are precisely what the naïve skeptic is being skeptical about, so any appeal to their normativity in order to ground their normativity will simply beg the question against this sort of skeptic. The beauty of Korsgaard's approach, and one of the reasons for adopting it, was that it promised to ground substantive moral claims in terms of the procedural normativity (the procedural normativity of action) that even this skeptic cannot wriggle out of.

This is perhaps why Korsgaard insists on the separateness of procedural and substantive considerations and claims to derive all of her conclusions merely from what is procedurally required in action. But now it is time to look carefully at what precisely is procedurally required in action, as described in the argument against particularistic willing.

Recall from Section I.iii that what distinguished an action from an event is that a certain procedure was followed in the case of an action, and if it was not the occurrence was merely an event. The procedure was simply this: the course of action taken was the causal result of some precedent-setting principle with which the agent identified herself. If the agent did not identify with some such principle in the action, it was not her action. If there was a principle present and causally responsible with which the agent failed to identify, then this is not a case of self-conscious causation, which is all that action is. If the agent failed to identify with a principle because none were present, then the agent should not be thought to be actively present in the action either, because there is nothing for her to identify with. Action is self-constitution because it is only when we act that we find something in the world that we identify as ourselves. Action constitutes the self because it provides something for the self to be, otherwise our selves are nowhere to be found. Action is what solves the mereological problem of how mere heaps of psychological urges become identifiable agents.

Korsgaard judges the Categorical Imperative to be the best principle of choice because it is best at achieving what is procedurally required of action, self-constitution. The Categorical Imperative tells us to choose the actions that can be universally endorsed by our future selves (and eventually, all other rational beings as well) and thus creates the most unified, best constituted selves possible.

But now look at the two different senses of "self-constitution" at play in the two paragraphs above. When we learned that all actions had to be "universal" rather than "particularistic" in the argument against particularistic willing, all this meant was that in

order to act, we must adopt some principle or other to identify with, otherwise we are nothing at all. Call this sense "self-constitution$_1$".

When it comes time to rank principles of choice like the principle of self-love, or the principle of consulting Emma against the Categorical Imperative, they are judged in terms of how well they create a stable, consistent, and coherent individual. The best principles are those that constitute the hardiest and most consistently identifiable characters. Call this sense "self-constitution$_2$".

Korsgaard readily admits that many different principles can be universal in the sense required in the argument against particularistic willing. Remember that "[s]omeone who takes 'I shall do the things I am inclined to do, whatever they might be' as his maxim has adopted a universal principle, not a particular one…"[76] Thus, any such principle, whatever its content, will be "universal" in the sense required by the argument against particularistic willing, and all of these principles will equally be cases of self-constitution$_1$.

Since all principles are equally universal in this sense of self-constitution, the only way they can be ranked (or even individuated) is in terms of some other feature. Korsgaard opts to rank them according to how well they succeed at the task of self-constitution$_2$, that is, the creation of a consistent and unified self. But this is precisely the sort of substantive value, external to what is procedurally required of action, that someone like the naïve skeptic can doubt. Railton's Gary might well reply, "Fine, I'm willing to grant that I must act on the basis of a principle. But I don't like consistent characters, they bore me. I'll constitute myself according to some other, more interesting principle." Korsgaard can show that this will lead to problems, unsuccessful projects,

[76] SC 27 and LL II 24.

disunity, and social strife. But all of these are substantive value claims that can be doubted by any naïve skeptic. We are condemned to act, not to successful careers, social harmony, or coherent life narratives.

In the absence of this substantive value, self-constitution$_2$, the procedural normativity only drives us to adopt some universal principle or other. It does not establish which principle of choice we must adopt. Therefore, it appears that Korsgaard cannot ground a single principle in terms of the procedural normativity of action. The practical necessity of her view underdetermines which principle we must adopt. It therefore does not ground moral normativity because it does not offer much practical guidance at all. Now, as we have seen in Chapter 1, I am committed to the idea that we could be legitimately guided by a principle against which all behavior accords, but if this is all Korsgaard has managed to establish, she will have fallen far short of her ambition of convincing the moral skeptic. She can still categorically insist, however, that agents morally ought to do anything at all.

VII. Mereology and Underdetermination

For Korsgaard, for there to be an agent at all, the mereological problem must be solved — that is that the heap of psychological constituents must have been made up into an agent — and it is that agent who acts.

In one sense, given this definition, action indeed is self-constitution, or it at least presupposes it, because action presupposes agents.

But given that there is no way to act without solving the mereological problem (for action requires an agent) this does not thereby also make *normative* the goal of providing the most long-lasting solution to the mereological problem.

Korsgaard thinks that actions that have the Categorical Imperative as their principles are the best actions because they are best at providing a long-lasting solution to the mereological problem. Korsgaard believes that Kantian sages are better than, say, Utilitarian maximizers, because they are more likely to survive *as agents*, sages will face fewer (if any) breakdowns or crises of agency.

This claim itself is questionable even if we grant that what is normative for agency itself is stability — but what is most crucial to notice for present purposes here is the fact that *stability* is not what is required of agency — only a solution to the mereological problem is.

The mistake is in the movement from the claim that in order to act you must solve the mereological problem to the insistence that action itself is *aimed at* the hardiest solution to the mereological problem.

If we grant Korsgaard's definition of action it is indeed senseless to ask "Why must I unify myself to perform an action?" because action implies an agent and an agent is thought to be something beyond a "mere heap" of psychological parts. But it does still make sense to ask "Why should I unify myself according to the Categorical Imperative, for other principles, such as the principle of utility, or the principle of immediate hedonism, will solve the mereological problem just as well?"

By claiming that action is both a solution to the mereological problem as well as *aimed at* the hardiest solution to the mereological problem, Korsgaard is able to claim for

her view satisfaction of both the universality constraint as well as the fallibility constraint. The satisfaction of the categoricity requirement comes from the fact that we can know, *a priori*, that every agent will be in the business of solving the mereological problem because that is what they must have done in order to *be* an agent, or in order to have performed an action. So, solving the mereological problem is what is constitutive of, and perhaps normative for, agency.

In that sense, it will apply categorically to all agents in that they must unify themselves to act. It makes skepticism about agential unification impossible. An agent can pose the question "Why must I unify myself in order to act?" but she is bound to fail if she attempts the skeptical feat of acting without unification. This is impossible, and in a way, patently absurd, because the heap cannot ask the question or perform the action —only an agent can.

But it is not quite right to say that every agent is necessarily "in the business of" solving the mereological problem, because many agents will have solved it in the past, in order to move on to other things, and the solution will be a byproduct. Other agents will have solved the problem equally well with solutions deliberately designed to have a brief lifespan.

In these cases it seems that we can quite sensibly claim that the various principles are capable of solving the mereological problem in better or worse ways. If the Categorical Imperative really did solve the mereological problem in a permanent and crisis-free way, we might rightly say that it is a *better* solution to the mereological problem than the Principle of Utility, provided that unification according to it left us open to the risk of dissolution back into a mere heap of parts.

In this way, it is again clear how Korsgaard can take her account to also meet the fallibility constraint, because there is a coherent and credible way here to judge and rank actions as better or worse. All actions must solve the mereological problem, these actions give longer-term solutions to the mereological problem, therefore, they are better solutions and are thus better actions.

The problem is, as we have seen, the feature of the account that meets the categoricity requirement and responds to the normative force will preclude a solution to the fallibility constraint and block the flow of normative force. Conversely, the feature of the account that allows for a coherent satisfaction of the fallibility constraint will not apply categorically to all actions, nor will it be guided by the normative force supplied by the necessity of action.

This is why: What is categorically (even perhaps analytically) required of an action for Korsgaard is that they come from an agent, and thus be the product of a mereological whole. To attempt to defy this is doomed to failure, because any behavior from a mere heap of psychological forces will be no action at all.

But *all* that is required of something to be an action is *that* the mereological problem be solved, and any principle is as good as any other at achieving this. An agent choosing according to the Categorical Imperative has solved the mereological problem as much as has the agent who chooses according to the Principle of Utility or an agent who chooses according to the Principle of Doing Whatever Upsets One's Mother. We know that this *must* be true because if the agent who chooses on the Principle of Utility had not solved the mereological problem by so choosing, there would be no agent there for us to consider.

So, in a sense, this normative requirement is universal and undeniable. All actions must solve the mereological problem because they cannot be actions unless they have. If an act is the mere behavior of an unorganized and disunified heap of psychological urges it simply is not an action. For much the same reason, it simply does not make sense to be skeptical about the necessity of solving the mereological problem because that is what is required of action itself. An agent might try to act without solving the mereological problem, that is to have an action emanate from a mere heap, but this will be impossible. (This is precisely what Korgaard uses the argument against particularistic willing to show.)

Here, the normativity of unification for action derives directly from the inescapability of unification for action. For Korsgaard, there is no way for an action to escape the normative force of the requirement to unify, because in order to *be* an action, this requirement must be met.

But the problem here is that *once* the action has occurred (or at least been willed) the mereological problem has been solved and the normative force can require no more. Unification is normative for action because action is impossible without it, but once an agent has performed an action, however silly, degenerate, or self-destructive that action may be, the normative demands of unification have been met, because there is an agent we can identify who is over and above the mere heap of psychological urges.

This feature of Korsgaard's account does indeed apply categorically to all actions and would indeed provide genuine normative force (assuming that the rest of her account is correct). But the fact that it applies categorically to all action is precisely what prevents the fallibility constraint from being satisfied. What unification demands is that

the mereological problem be solved either before, or in the process of, action. And because we can recognize the utility-maximizing action as an action (i.e. as something that emanates from an organized whole creature aimed at maximizing) we can infer that the mereological problem has been solved.

To think of the mereological problem as one that can be solved is apt, because it, just like a mathematical problem or a logic puzzle, can stand either as solved or still unresolved. It cannot be solved in matters of degree. The salient question when considering the mereological problem is "Is there an agent there or not?" and this is not a matter that admits of degree. Even though it refers to a "heap" of psychological urges, there is no sorites paradox or vagueness about when there is a heap as opposed to an agent for Korsgaard. If there is a unifying principle in the explanation of the action, there is no longer merely a heap. If there is no unifying principle, then there is nothing more than a heap. The mereological problem is one that is either solved or it isn't for any given behavior from a psychological entity, and all that is normative for action is that the mereological problem be solved.

Given this, it is easy to see how in satisfying the categoricity requirement, Korsgaard's view is incapable of satisfying the fallibility constraint. This is because there can be no degenerate instances of this kind. An action demands that the mereological problem be solved, and any principle on which an agent acts will count as having solved it. Thus, any action that does not provide a solution to the mereological problem (and thus does not satisfy the normative requirement for action) will fail to be an action at all. Pseudoactions that emanate from mere heaps are themselves mere behavior, not actions.

Return now to a phrase I used earlier that "all actions are in the business of solving the mereological problem." We can see now the sense in which this is true, because all actions, in order to be actions, on Korsgaard's account, must have been in the business of solving the mereological problem to have become actions in the first place, but this does not ensure that all actions will be or must be *aimed at* the business of solving the mereological problem once they have become actions. An action grounded entirely in the Categorical Imperative provides just as much opportunity and ability to distinguish an agent in the heap of motivations as does an action specifically chosen according to a self-destructive principle. Because both of these actions successfully provide an agent, they are equally responsive to and successful in light of the normative force and requirement provided by the necessity of action.

A skeptical challenger risks incoherence if she attempts to act without solving the mereological problem, but there is no similar incoherence in a solution that neither strives to keep nor succeeds in keeping the mereological problem solved. We might join Korsgaard in criticizing the maxim that requires that we "Live fast, die young, and leave a good-looking corpse," but all of our criticisms will have to be "external" in the sense that Korsgaard's uses it here, i.e. not inside the requirements of action itself, but further, extra, and therefore optional requirements. There are many ways we can join Korsgaard in criticizing a principle that recommends a fast life and an early death, but Korsgaard herself cannot claim that such a principle generates degenerate actions because we are perfectly able to identify characters like James Dean above and beyond the heap of psychological urges that constituted him, and this is all that is necessary —both constitutively and normatively— of action on Korsgaard's account.

VIII. The Case of Constitutivism

So far, we have seen how Korsgaard's later moral philosophy in the Locke Lectures provides satisfactory answers to both the categoricity requirement and the fallibility constraints, just not answers that can be simultaneously satisfactory.

At this point we might wonder if this is a genuine idiosyncrasy of Korsgaard's view or if it is a more general problem. What would have happened, for example if Korsgaard had managed to ground the normativity of action in a requirement whose satisfaction really did admit of degree. Suppose instead that Korsgaard's claim bottomed out in the necessity of action and what was constitutive of action was that it *aim at* agential stability.

If such an account could be grounded, we would seem to have two options for understanding how the normativity could be grounded, neither of which seem to revive much hope for jointly satisfying the categoricity requirement and the fallibility constraint, nor in deriving normativity from constitutive standards.

The first option (which seems the most sensible, in this case) is to claim that while one can constitute a stable agent in ways that are better or worse, one cannot *have the aim* of constituting a stable agent in a greater or lesser degree. Thus, if we are to normatively evaluate actions in terms of how well they satisfy their constitutive standards, we will again face the impossibility of finding degenerate instances of a kind, since all actions will satisfy their constitutive aims perfectly, and all that do not will fail to be actions.

If we wish to avoid this conclusion and instead insist that having an aim is something that can be satisfied in a matter of degrees, we will face a new problem. Consider for example, an agent who only satisfies this constraint halfway: he only

halfway has the aim of creating a stable agent, and is thus only halfway an agent. Here, it seems that we lose the ability to engage in normative criticism or apply motivating normative force. Suppose we were to exhort him "Come on! You can do better than that! Unify yourself even more, that's what you're trying to do anyway!" It seems that he could, quite reasonably reply, "Why? I'm only attempting to be halfway an agent, and on that score I couldn't be doing better." Korsgaard claims in the Locke Lectures that every housebuilder is attempting to build a good house, even the best house, in the very act of housebuilding.[77] But it is far from obvious that this is true. I can set out to build a merely cheap and adequate house perfectly comprehensibly. Further, I can quite easily deflect the potential criticisms that my roof leaks and my door squeaks by pointing out that I only *partly* had the goal of building a good house, and on that score, I've succeeded completely. And it seems that this sort of immunity to the normative force for improvement will be available to any sort of item or activity if we believe that kind membership deteriorates along with normative success. If a house is less of a house because it achieves the goal of a house less well than it might, it becomes difficult to see why it must necessarily strive to be a perfect house. If it only partially has the goal of being a house (or a good house), it has achieved this goal once it becomes a partial (or partially good) house. There is no normative force left to push it toward being a better house, if we allow that it only partially had the goal of being a house in the first place.

The puppet Pinocchio just so happened to fully have the goal of being a real little boy, but if he'd only had the goal partially he could have settled on being the reasonable approximation that he was without necessarily feeling any normative push to be more.

[77] See Locke Lectures I and II.

Since neither Korsgaard's own account, nor either of the further options gestured at here seem to have much hope of jointly satisfying the categoricity requirement and the fallibility constraint, it looks like there is reason to doubt that any constitutive account could. If this is correct, the tension between these two requirements may extend beyond Korsgaard's particular account and into any constitutivist project.[78] This is a point I will return to in Chapter 5.

[78] Matty Silverstein insists to me that the constitutivist account he gives in his dissertation does not suffer from this problem, but I have not yet had the chance to evaluate this claim.

Chapter 4

Butler, Brutes, and Bad Action: A Case Study from the Past

I. Introduction

Joseph Butler's ethics clearly stand in the same general autonomist tradition as do the ethical systems of other autonomists like Kant and Korsgaard who would follow him, and whose views I have already examined in the previous chapters.

Butler is an insightful and influential historical antecedent to these two currently more visible philosophers broadly in his tradition, so his view presents itself to examine in terms of how it might explain bad action and thereby accommodate the fallibility constraint. If his account is found to suffer from a similar difficulty, this will provide a further piece of data to suggest that perhaps such a problem is symptomatic of the entire family of views in this tradition. If his system can instead avoid this problem and account for bad actions, perhaps this will provide a solution that can be translated into terms that could aid the other theories in his tradition. As it happens, Butler's account is able to accommodate the fallibility constraint in a way that provides a lesson in what to avoid if an account is to be able to satisfy the fallibility constraint.

With this in mind, I will examine in this chapter if and how Butler's ethics and moral psychology can account for bad actions, and what the implications are for both his own and other theories. To do this, I will present a chapter in seven sections, including

this introductory section. First, I will briefly revisit the sorts of problems I take Kant and Korsgaard to face, so as to provide cautionary tales about what Butler's account cannot do if it is to account for bad action. Next, in order to understand and contrast how things might go wrong in the case of bad action, I will have to explain how things can go *right* and provide Butler's account of morally good action. Third, and against this backdrop, I will show how Butler's account generally allows agents to act viciously in a way that satisfies the fallibility constraint. Fourth, I will detail the three ways in which an action may be vicious, and show how the third way poses a serious problem if a general *obligation* to virtue is to remain a coherent notion. Fifth, I will suggest that while this difficulty may remain a difficulty in theory, it can be (and often is) practically solved through moral education. Finally, I will briefly conclude that while Butler's solution to the bad action problem is a viable way to satisfy the fallibility constraint, it is doubtful that his success can translated into terms amenable to the Kantian or Korsgaardian projects. In Chapter 5 I will address the question of whether the strengths of Butler's solution can translate to other moral theories.

II. Kant and Korsgaard

Apparently, Kant's and Korsgaard's accounts leave no room for bad action because they both take the constitutive standards of action to determine the moral value of those actions. The feature that makes an event an action and the feature that confers on that action the status of moral goodness is the same thing for both Kant and Korsgaard. Consequently, actions and goodness will stand and fall together. Properly considered, all actions are good actions; all instances of moral vice will be instances where agency is absent or its power usurped.

i) Kant

According to commentators at least dating back to Henry Sidgwick in 1888,[79] Kant has problems accounting for bad actions because he often claims both that an agent's actions are only those that she performs autonomously, and that morally good actions are just autonomous actions. From these two premises, it follows that the only actions are morally good actions. All other apparent actions are mere behaviors that are the causal result of something (a desire, inclination, previous phenomenon) that is not the agent. This scheme clearly leaves no room for moral responsibility, since every instance of bad "action" is just an instance where the agent was not in control, and thus was not free to have done otherwise.[80] But Kant believes that agents are free (at least noumenally) to do otherwise, and wishes to hold them morally accountable even for their misdeeds.[81]

Sidgwick himself believes the tension between these claims comes from the fact that Kant uses the same term, "freedom," to refer to two quite different phenomena. One kind of freedom, the kind that Sidgwick calls "moral" or "neutral" freedom, is just the capacity to choose good or ill in any given circumstance.[82] The other kind, which Sidgwick calls "rational" or "good" freedom, is the kind an agent displays in a fully autonomous and morally good action that is in accordance with and comes from respect for the moral law. The first is noumenal freedom from antecedent phenomenal causes; the second is autonomous freedom from foreign and morally bad influences.

---

[79] Sidgwick, 1888.
[80] This way of reading Kant is admittedly controversial, but as I am using it as a springboard to motivate concerns in Butler's ethics, I am not primarily concerned if it is ultimately the right reading of Kant. I here appeal to Sidgwick's account of the problem as a famous and insightful formulation of it.
[81] As I claimed in Chapter 2, freedom to do otherwise is not what concerns me, but it is what concerns Sidgwick, so I mention it here.
[82] Notice here that the freedom to do otherwise and the freedom to go wrong are conflated.

The two are both freedom, but they are not one. Kant's theory faces problems whichever one he might settle on. He will have to sacrifice either the moral normativity or the moral responsibility thought to stem from freedom. Should he ultimately claim that freedom is the neutral freedom of a noumenally free agent, he will lose his, as Sidgwick puts it,

> "spirit-stirring appeal to the sentiment of Liberty…as idle rhetoric. For the life of the saint must be as much subject…to the necessary physical causation as the life of the scoundrel: and the scoundrel must exhibit and express his characteristic self-hood in his transcendental choice of a bad life, as much as the saint does in his transcendental choice of a good one." (KFW 516)

By this Sidgwick means that should Kant opt for this notion of freedom, he will have to sacrifice his most appealing and characteristic (Sidgwick says "fascinating") ethical idea "that a man realizes the aim of his true self when he obeys the moral law." (KFW 516)

If Kant instead decides to retain the "fascinating" freedom that is the rational freedom of an autonomous agent, "the whole Kantian view of the relation of the noumenon to the empirical character will have to be dropped, and with it must go the whole Kantian method of maintaining moral responsibility…" (KFW 516) Notice that Sidgwick's construction of this problem has higher stakes than most. Most versions of this problem only work to show that an agent is not responsible for her *bad* actions, and so while she cannot be held accountable to moral sanction, she still may be praised for the good actions that are truly her own.[83] As Sidgwick sets up the problem, if "freedom" is just the rational freedom of an autonomous agent whose will is subject to moral laws, then it is *not* the noumenal freedom from phenomenal necessity used to overcome determinism. This will mean that an agent is not responsible for *any* of her actions, good

[83] This is reminiscent of the way Susan Wolf would have us understand things to be in her "Asymmetrical Freedom." See Wolf, 1980.

or bad, subject to moral laws or not. As Sidgwick reads him, Kant can either have noumenally free actions that are either good or bad, or no actions at all.[84]

ii) Korsgaard

Sidgwick traces Kant's problem to an equivocation on "freedom." Korsgaard's view faces a similar problem, except her equivocation is on the term "universal." I have already detailed the difficulties that face Korsgaard's view in the preceding chapter, but Sidgwick's framework can be used to show that the problem in Korsgaard's account is analogous to the one in Kant's, bearing a sort of family resemblance to it. As we have seen, in the Locke Lectures, she claims that "action is self-constitution" and appeals to the "necessity of action" in order to ground her notion of moral normativity. She intends to make this normativity extend to all agents, even those who do not recognize its authority or are not sensitive to it. In order to act, says Korsgaard, we must constitute ourselves by allying our selves with a principle of choice. To do this we must identify with some "universal" principle; that is, some principle that speaks in universal terms and is taken to set precedents for future cases. In order to distinguish good from bad actions, we may rank them according to how "universal" they are; that is, how well they constitute a stable and unified agent. Since every action aspires to this universality, and this universality will imply morality, every agent is in fact committed to morality, even if the agent does not yet notice this fact.

[84] Of course, this is not quite how I wish to understand things, since I take the Strawsonian reactive attitudes to provide a basis for moral responsibility that can survive the truth of determinism, provided that mental content can also survive that determinism. Once again, in any event, I only intend here to present Sidgwick's understanding of things, not my own.

The deep problem with this view is that the sense of "universal" required of a principle for it to generate actions is equally true of all principles, and it is only this weak sort of universality to which skeptics are committed. This sense of universality cannot distinguish good from bad actions, since it is equally present in all actions.[85] The way Korsgaard is actually able to rank the various principles is in terms of another value, agent unity, but this is something that skeptics can easily reject, since not even Korsgaard thinks agency is necessarily committed to this.[86] On her account, the feature of action that is meant to ground normativity is not a feature that can distinguish one action from another; the feature in terms of which actions may be morally ranked is not normative.[87]

Both views face problems from joining normative and constitutive standards for action, reflected in a terminological confusion. As it happens, few of these problems (especially those that Kant suffered) will be overt in Butler's moral psychology. But, by way of foreshadowing, notice that many of Korsgaard's problems come from attempting to get moral normativity to extend over even those who do not agree with or are not sensitive to its claims. A feature of Butler's account may share precisely this problem in this respect, if Butler is determined to share Korsgaard's ambitions.

---

[85] Just as "neutral" or "moral" freedom is equally present in all noumenally free actions on Sidgwick's presentation of Kant's account.

[86] I should be careful here since Korsgaard does say things like "I also believe it is essential to the concept of agency that an agent be unified." (LL I 15) As I argued in Chapter 3, this means that an agent, for Korsgaard, must have gotten itself unified in order to exist, not that it continues to be committed to the business of maintaining this unity.

[87] Just as on Sigdwick's version of Kant, the feature of actions that allowed an agent to have chosen otherwise does not give them moral worth, but the feature that gives them moral worth does not allow them to have chosen otherwise.

III. Butler's Account of Virtuous Action

Most of Butler's view concerning what is required for cases of virtuous action is summarized in a statement he makes in the Preface to his *Sermons*.[88] A brief analysis of this remark should give a good outline of Butler's account of virtuous actions. In the Preface, he states

> [I]n reality the very constitution of our nature requires that we bring our whole conduct before this superior faculty, wait its determination, enforce upon ourselves its authority, and make it the business of our lives, as it is absolutely the whole business of a moral agent to conform ourselves to it. (S.Preface.25)

The "superior faculty" Butler mentions is the faculty he interchangeably calls "conscience" or "the principle of reflection." It is the faculty "in man by which he approves or disapproves his heart, temper, and actions." (S.I.8) God has designed this faculty so as to always[89] give the correct answer concerning which action morally ought to be performed. Further, he has granted the principle of reflection the ability to restrain agents from performing even those actions toward which they have the strongest desires. This is because God has designed the principle of reflection to be different from the other principles that recommend action "[a]nd this difference, not being a difference in strength or degree, [Butler calls] a difference in *nature* and in *kind*." (S.II.11) Because we can both find it in our constitution and recognize the supreme position it was designed to have in that constitution, we may infer that our proper natural function (i.e. our virtuous action) consists in our respecting its supremacy among the principles that guide action, and thereby continually consulting it and abiding by its dictates.[90]

---

[88]Butler, 1983. Passages in the *Sermons* will be referred to by sermon and paragraph number, prefaced by S; those from the *Dissertation* by paragraph number prefaced by DV.
[89] When functioning properly and not interfered with by external forces.
[90] Note the teleology in these claims. This is a matter to which I will return in Chapter 5.

Given this picture of the resources available in human moral psychology and their relations to one another, it should now be clear why "the whole business of a moral agent" is to learn and then follow the dictates of conscience. Though Butler lists three requirements in the passage quoted above, this scheme of moral agency really only provides two necessary criteria for virtuous action. For an action to be a virtuous one, an agent must only 1) consult conscience about what is to be done, and, then, 2) respect conscience's answer to this question by doing as conscience dictates. The stipulation that an agent wait to learn what conscience dictates is superfluous if both of these criteria are met, since one will necessarily have to have waited to learn conscience's answer before one respects its recommendation by acting accordingly.

Even if superfluous, this criterion is still instructive in highlighting the necessary *order* in which these criteria must be met in order to genuinely be case of virtue. Requiring that an agent first consult conscience and then act accordingly explicitly excludes from virtue cases wherein an agent merely acts on the basis of some principle other than conscience, and then seeks retroactive vindication by pointing out that her act was what conscience would have dictated anyway. Virtue cannot be gained in this manner. It is a matter of cool antecedent reflection, not of fevered rationalization after the fact.[91]

IV. Bad Actions on Butler's Account

i) Capacity and Responsibility

[91] Notice that this does not require that virtuous actions must be *immediately* preceded by conscientious reflection, just that the relevant reflecting has been at some point prior to the action. As we will see in the upcoming subsection iii, conscience can sometimes allow (and even require) that one *not* reflect just before acting.

Now that we have seen Butler's account of how agents can get moral agency right, it remains to be seen if he has a viable[92] account of how agents can go wrong. Kant ran afoul of bad action by being committed to the idea that those who acted viciously were ultimately not responsible for their wrongdoing, since vice cannot be freely (i.e. rationally) chosen. In order to avoid this unsavory outcome, Butler's account of agents will have to be such that they are capable of freely making the wrong choice so that they can be held responsible for it. This freedom must be the "neutral freedom" Sidgwick discussed, the freedom to act viciously or not; otherwise there is no room for moral responsibility.

Butler himself notices (among other things) exactly this point in the *Dissertation* when says

> We never, in the moral way, applaud or blame either ourselves or others for what we enjoy or suffer, or for having impressions made upon us, which we consider as altogether out of our power, but only for what we do or would have done had it been in our power; or for what we leave undone, which we might have done or would have left undone, though we could have done it. (DV.2)

This suggests a generally counterfactual account of moral responsibility. We may only sensibly "applaud or blame" an agent "in the moral way" if she could have done otherwise than she did, or would have done otherwise than she did, had she been able.[93] Though Butler's language emphasizes action, it is clear what he is concerned with is intention. Here, Butler anticipates Strawson and Frankfurt, just as he anticipated Kant and Korsgaard before. It is not what an agent actually accomplishes that primarily

[92] It seems appropriate here to note that I am taking an account of bad action to be "viable" if and only if it allows one to simultaneously do wrong and yet remain an agent. Only agents are responsible for their actions, and if no agent can also be a wrongdoer, there can be no conceptual room left for moral responsibility.

[93] I do not necessarily agree with Butler here, since the ability to do otherwise does not entail the ability to go wrong, but I believe this is an accurate representation of Butler's account.

matters for moral responsibility, but rather what the agent wills, and thus the actions (of commission or omission) that do follow if the agent has sufficient power, or would follow if the agent did.

Butler also clearly believes that we are responsible for our wrongdoing. In the preface to the *Sermons* he says "Our constitution is put in our own power. We are charged with it; and therefore we are accountable for any disorder or violation of it." (S.Preface.14) Beyond any of the reasons I provided in support of the fallibility constraint in Chapter 2, Butler is, by his own constraints, committed to providing a viable account of bad action. If we are to be accountable for disorders of our constitution,[94] and the vicious acts that result from this disorder, we must be capable of having ordered ourselves when we did not. Otherwise, we cannot be morally accountable for this disorder and the vice that it constitutes.

Butler does not merely insist that moral responsibility requires this sort of capacity; he meets his own constraints by providing a moral psychology that coherently accounts for how such freedom is possible and in what it consists. Beyond this, it appears that his moral psychology also meets *my* constraints for viability, by conceptually allowing agents to misbehave, and so satisfying the fallibility constraint.

Perhaps Butler is able to provide a viable moral psychology and thereby avoid the mistake that Kant and Korsgaard commit because he is particularly aware of it. The mistake is making the feature that gives actions their moral worth apply equally to all genuine actions. This makes the feature trivial, and thus no guide in choosing one action

[94] These "disorders of constitution" will always be cases in which we do not conform ourselves to conscience. It is telling that Butler thinks we can be responsible for our constitution even when it is "disordered."

over another.[95] Butler is clearly aware of the dangers of this move in his second sermon. There, as he argues against the claim that actions that come from any of our internal principles other than conscience are just as "natural" as those come from conscience itself,[96] he says

> If by following nature were meant only acting as we please, it would indeed be ridiculous to speak of nature as any guide in morals, nay, the very mention of deviating from nature would be absurd; and the mention of following it, when spoken by way of distinction, would absolutely have no meaning. For did ever any one act otherwise than as he pleased? (S.II.4)[97]

Here, Butler recognizes and avoids the very problem that plagues Kant and Korsgaard, by noticing that if morality is to be a guide to the free actions we choose and are held responsible for choosing, it cannot be present in all actions.[98] Given the data we have from our reactive attitudes, there must be a difference between moral and immoral actions on the basis of which we can choose, and on the basis of which we can rightly resent or admire others for so choosing.

To show that Butler's account does indeed offer all it promises, I will argue as follows. First, in the next subsection, I will show that Butler's account meets his own constraints by explaining his moral psychology. I will then show that it meets my constraints by detailing the three ways in which vice is possible on Butler's account in Section V. This will not be enough, since the discussion in Section V will show that the

---

[95] Of course, I will have no complaint against the idea that such a feature might be the basis for a perfectly legitimate and normative rule, it is only that it cannot be the basis for our morality, since we know, from our reactive attitudes and elsewhere, that we have evil and ill-will to separate from goodness and benevolence.

[96] After all, they come from principles with which we are naturally endowed.

[97] Again, as I claimed in Chapter 1, there probably is a sense in which one can "act otherwise than as he pleased" but the crucial point here is that a criterion shared equally by all potential actions is no ground for separating the good ones from the bad.

[98] That is to say it cannot be present in all actions as a necessary conceptual truth. This does not exclude the possibility (however unlikely it may be) that one might contingently perform only morally good actions.

third way in which vice is possible generates a puzzle. This puzzle will be discussed and largely resolved in Section VI.

ii) Agents and Brutes

Butler is able to both allow for bad actions and hold agents morally responsible for those actions because he distinguishes between actions and events not in terms of the *circumstances in which* they come about,[99] but in terms of the *being from which* they come about. In Butler's terms, actions are the things that agents do; all else that might appear to be action is merely the behavior of brutes.[100] The thing that distinguishes an agent from a brute is not behavior; it is the extent of their respective mental faculties and their capacities to use them. Brutes are moved by passions, instincts, and principles like the principle of self-love. Such is the case for agents as well, except they also have conscience, the capacity to reflect, and this is what makes them agents. Butler draws and employs this distinction in numerous places, but it is perhaps most vivid in the opening sentences of the *Dissertation*.

> That which renders beings capable of moral government is their having a moral nature and moral faculties of perception and of action. Brute creatures are impressed and actuated by various instincts and propensions; so also are we. But additional to this, we have a capacity of reflecting upon actions and characters, and making them an object to our thought… (DV.1)

The sense of being free to do otherwise that is relevant on Butler's account is not relative to the given behavior, but to the creature behaving. When heteronomy befell agents on the Kantian conception, it is not clear that they could have done otherwise in those

[99] As Kant and Korsgaard do, claiming that actions only occur when the proper procedures are followed.
[100] This is not to say that on Butler's view even an agent's sneezes and seizures must count as actions —in this way Butlerian agents are able to merely behave— but any deliberative action will be properly subject to an agent's conscience. Thanks to Sean Kelsey for helping me see this.

circumstances. The point is not that the heteronomous "agent" was necessarily incapable of pursuing another course of action, it is that there was *no category* for the heteronomous "agent" to fall into and count as morally worthy yet still heteronomous. Here, the immediate circumstances of the action determine its moral worth. Conversely, when vice plagues Butlerian agents, they can be held responsible for whatever they do, since they are the *kind* of creatures who are capable of reflecting on their behavior and therefore have the capacity to misbehave even in the full flush of their agency. To learn if the behavior of a given creature is an action in Butler's sense, we do not need to know if conscience was actually playing a part in the given behavior. We simply need to determine whether the creature in question had conscience as a part of its psychological repertoire and thus, in principle, could have appealed to it as it should have done. Thus what makes something an action on Butler's account (i.e. that it was done by a creature with a conscience) is separate from what makes that action morally worthy (i.e. that conscience was actually consulted and obeyed). It is important to note here that the point is *not* that the creature must have been metaphysically able to do otherwise than it did (we have seen this is no ground for the fallibility constraint), instead, it must be capable of *being* otherwise than it was. It must be able to survive through its choices, be they morally worthy or not. What is important is that, even when misbehaving, the agent not cease to be the sort of creature that it is and which thereby makes it subject to moral demands in the first place. It is easy to see what makes a bad action for Butler: it is when an agent does not use the conscience God gave him.

Butler claims elsewhere in the *Dissertation* that this is in fact how we determine moral responsibility, by comparing the behavior with the creature performing it.

> [O]ur perception of vice and ill desert arises from, and is the result of, a comparison of actions with the nature and capacities of the agent.[101] For the mere neglect of doing what we ought to do would, in many cases, be determined by all men to be in the highest degree vicious. And his determination must arise from such comparison, and be the result of it; because such neglect would not be vicious in creatures of other natures and capacities, as brutes. (DV.5)

As in other theories, precisely the same behavior might be morally weighted or morally neutral, but what is unique to Butler's account is that he does not draw this distinction in terms of whether or not the proper procedures were followed; but only if the behavior was that of a creature *capable of following* the proper procedures.

Given his conception of agency, it is easy to see that Butler gives us a viable account of bad action. There is no contradiction involved in understanding what happens when an agent performs a bad action for which she is morally responsible. All that is required is that she be a creature with a conscience (thus she is an agent), and that in a given instance she fails either to consult or to abide by her conscience (thus she is vicious). Because she had a conscience available to her appeal, she can be held responsible for failing to appeal to it.

iii) Forgiving Failures of Reflection

In order to avoid some confusion before providing a taxonomy of vice, it is worth a moment to notice precisely what will count as a bad action on Butler's account. The

[101] It is no trouble at all that Butler here refers to all behaving creatures as "agents" performing "actions". In light of other passages, this must just be an instance of Butler speaking loosely. He himself draws the distinction between action and event at DV.2, he there makes it clear that brutes are not beings who perform actions, and thus are not agents: "It does not appear that brutes have the least reflex sense of actions, as distinguished from events, or that will and design which constitute the very nature of actions as such, are at all an object of their perception." (DV.2)

above implies that Butler will allow no excuse to forgive an agent's failure to reflect prior to acting; and this is, in a carefully qualified sense, precisely right.

Conscientious action is, for Butler, "absolutely the whole business of a moral agent" (S.Preface.25). Though he puts the words in an objector's voice, it is clear that he endorses the notion that virtue and religion require "that the *whole* character be formed upon thought and reflection, that *every* action be directed by some determinate rule, some other rule than the strength and prevalency of any principle or passion." (S.II.2, emphasis not added)[102] Butler clearly has rigorous requirements for the exercise of virtue, and simply does not allow for extenuating circumstances to forgive failures of reflection. And this is hardly surprising, since Butler gains the capacity to viably account for bad actions by comparing behavior to beings behaving, not behavior to the circumstances of behaving. If he were to forgive agents for failing to reflect in circumstances where they ought to, he would be shifting to a circumstance-based evaluation of moral goodness, and thereby substantially abandoning what is a characteristic strength of his account.[103] This requirement is strict, and it does not appear that Butler can abandon it without great cost, but we should notice that this does not hobble his account in a way one might initially think.

Requiring that "*every* action be directed by" conscience might look to constrain proper moral deliberation to a process that is so cumbersome as to be useless. If this requirement amounts to demanding that every action be immediately preceded by a long and careful exercise in conscientious reflection, then Butler's account is probably

[102] Butler's anticipation of Kant and Korsgaard is especially vivid here, in his emphasis on the importance of rules over the mere strengths of particular passions.
[103] I should be careful to note here that, as I will claim in Sections V and VI, there are some very special epistemic circumstances that will require that we forgive agents for not reflecting. But I will claim that this is because these circumstances will serve to alter the sort of being one can function as.

doomed. There are numerous cases in which what is morally required is expeditious action, not a pause to reflect. Imagine a case a simple as catching someone as he falls. In order to actually succeed in this task, you must act almost instantly and reflexively, not stop to consider whether or not you should reach out to catch him. If Butler's account is to provide a genuine alternative to Kant's or Korsgaard's, it will have to be able to accommodate a requirement as elementary as this.

Happily enough for Butler, his account can do so easily. Butler himself pointed out "the paradox of egoism," which is just the observation that often one's happiness is best pursued by not attending to its promotion at every moment. As Stephen Darwall has pointed out,[104] precisely the same sort of observation can be made on behalf of what conscience requires. In what he gently refers to as a "paradox of conscience", he notes that

> A person may be likelier on certain occasions, in certain areas of life, and in the long run, to act in ways he would approve of on reflection "in a cool hour" if he were not explicitly to deliberate and form a judgment before each action. … The person who is guided by his best judgment is not explicitly guided act by act. (SD 422-423)

Darwall's observation is straightforward enough: conscientious reflection dictates what is morally appropriate in all circumstances, and one of the things it can dictate is that there are circumstances in which it is morally appropriate to cease reflection or skip it altogether.

But notice that this does not amount to *forgiving* agents for failing to reflect when they ought; it is just to notice that there are certain circumstances in which they *ought not*

[104] Darwall, 1988. Hereafter referred to in the text as SD with page numbers.

reflect.[105] This does not alter in the least Butler's requirement that the actions that are genuinely good are those that come as a result of (and therefore after) conscientious reflection; it just points out that such actions can still be good if they come substantially later than the reflection that recommends them. In order for an action that does not immediately follow reflection to count as morally good, its agent must cease or skip reflection because she recognizes her circumstances as those in which conscience demands that she cease or skip reflection. And she can only be capable of such recognition if she has antecedently reflected on circumstances such as these. Any other case in which she acts in accordance with what reflection would have dictated without the relevant previous reflection will not be a case of morally laudable action, but a case of fortunate reflex. Notice here the circumstances of the bad action are still morally *relevant* but in such a way as to salvage an agent's responsibility for it, rather than eliminate it. Yet again, an action will not count as morally good even if it can be shown that it was the act that conscience would have dictated, had one consulted it. It must be that act motivated by the fact it is what conscience dictates; only then is it the act*ion* that conscience demands and thus deserves moral praise.

For these reasons we should insist that Butler's account has no room to excuse agents from failing to reflect when they ought. But this does not mean that it cannot make room for (and even recommend) absences of reflection altogether.

[105] This is not to imply that Butler has place for morally neutral actions. Conscience might also point out circumstances that are morally innocuous and do not require further moral consideration. But we are only justified in treating morally neutral circumstances as such if we conscientiously recognize them as such.

V. Three Flavors of Vice

In Section III, I claimed that there are only two (ordered) criteria for virtuous action on Butler's account. It is natural to suppose in light of this that there are exactly two ways to be vicious, either by failing to consult conscience or by failing to abide by its dictates. Broadly, this is correct; but there is a salient distinction between the two ways one can fail to consult conscience, and it will be important to consider these separately. One way will be unproblematic, in that it already acknowledges the authority of conscience and recognizes "our obligations to the practice of virtue". (S.Preface.12) The second way is more troublesome, and it threatens to question the authority of conscience by remaining ignorant of it, and remain in territory to which obligations seemingly cannot extend. Consequently, I will here be discussing three distinct ways in which an agent may be vicious on Butler's account, in order of least to most troublesome for his theory.

The first way in which an agent may be vicious is, temporally, the last. Instances of this sort of vice occur when an agent does bring her conduct before conscience, reflects carefully and thoroughly about what she should do, comes to the correct conclusion, and yet fails to abide by what her conscience dictates. It is probably surprising that I am ranking this sort of vice as the *least* troublesome for Butler's account, since it sounds most like the worry that has bewitched the internalists and externalists of recent moral philosophy. But it is no special problem for Butler's moral philosophy, because any way in which such vice might occur involves an agent who already recognizes the authority of conscience. Whether the agent willfully acts against conscience's dictates or attempts to abide by them and fails, the agent knows that conscience is authoritative; she is just unwilling or unable to abide by its demands. In

either case she is vicious and knows she ought to do otherwise. Such a case is also clearly not in violation of the fallibility constraint. Here, we have a full and conscious agent who nevertheless counts as having failed to do as conscience dictates, and so counts as having acted, but badly.

It is no problem that she *can* will to do other than what conscience dictates, because she knows (when she reflects or appeals to conscience) that she *ought* to do as conscience dictates, given the sort of creature she is and was designed to be. It will come as no surprise to her if or when she is held morally responsible for acting against her conscience. After all, she knew better.

Equally unproblematic for conscience's authority are cases in which weakness of will or weakness of body impede an agent from following through with conscience's dictates. Here again the agent knows what ought to be done and that it ought to be done. It is only mental or physical incontinence that keeps the agent from doing what ought to be done. It is clear what went wrong, and what must be changed. The agent must strengthen herself (mentally and physically) so that she will be able to do what conscience dictates of her.[106]

One general way to be vicious is to fail to follow the dictates of conscience once it has been consulted; the other is to fail follow the dictates of conscience because it never was consulted. Remember it is not enough to act as conscience would have dictated, had one consulted it. Agents can only be virtuous if they do as conscience dictates *because it is what conscience dictates*. Equally, an action can only be the right one if it is what

[106] Butler can likely forgive failures to act that come from weakness that are such as to put such actions "altogether out of our power", but it is not clear how far such forgiveness extends. We might argue at great length about how far one's obligation to develop one's capacities extends, but this is not relevant for present purposes. None of these cases question the fundamental authority of conscience, just the extent of one of its dictates.

conscience dictates and is done *because* it is what conscience dictates. This sort of vice may or may not threaten the authority of conscience, depending on the circumstances of the agent who fails to reflect.

For an agent who is accustomed to (or even just acquainted with) reflection, the failure to reflect can still sensibly be thought to count as vice, as shirking an obligation, without threatening the authority of conscience to oblige agents to consult and obey it.

Often in the discussion so far, I have referred to conscience as if it is were like an appliance that can be turned on or off, implying that the agent might passively await conscience to give an oracular response that the agent does not and need not understand. But this is to seriously mischaracterize conscience, which is also tellingly called "the principle of reflection." Consulting conscience should properly be thought of as a process of careful consideration and reflection that the agent engages in and maintains. Thus, when conscience concludes its considering and recommends a course of action, the agent will fully understand why this is the right course of action, since she has been reflecting on it all along.

I bring out this distinction now to point out two different ways in which an agent who is already acquainted with conscience might either fail to consult it or fail to consult it adequately. Once an agent is already acquainted with the power and authority of conscience, she might fear that its dictates would restrict her preferred course of action and thus not consult it at all, or she might also cease or sabotage reflection once she suspects her reflection will lead her to a conclusion she dislikes. Butler is aware of both of these techniques, and (quite rightly) classifies them both as vice. In his seventh sermon he notes:

> [I]f there be any such thing in mankind, as putting half-deceits upon themselves; which there plainly is, either by avoiding reflection, or (if they do reflect) by religious equivocation, subterfuges, and palliating matters to themselves; by these means conscience may be laid asleep, and they may go on in a course of wickedness with less disturbance. (S.VII.10)[107]

Notice that these processes allow us to "go on in a course of wickedness with less disturbance" they do not somehow excuse our vice or undo our obligations; they are simply techniques we can employ to more easily act against our consciences. Either way of avoiding conscience is clearly a violation of the duty that grips us as agents requiring "that we bring our whole conduct before" conscience. Any time we do not consult conscience before we act,[108] we violate this obligation.

Furthermore, though these strategies deliberately attempt to undermine conscience's authority, they both acknowledge conscience's power and authority in the very process of attempting to evade its dictates. Indeed, it only makes sense to avoid conscience's conclusions if one recognizes them to be authoritative and powerfully binding. If conscience could not have the final say about what one ought to do and could not exercise powerful (often decisive) motives to act, why not give it its turn to speak? Even if one suspects that conscience will advise against the course of action to which one is most partial, there seems little harm in consulting it if it can easily be ignored, if its voice is just one among many. It is only if an agent recognizes the authority that conscience has in presenting binding obligations that it makes sense to avoid or cut short the careful reflection that might exclude the presently desired course of action.

---

[107] Butler, 1873, p. 88. References to sermons outside the five published in the Darwall edition will follow the same convention of being referred to by sermon and paragraph number, prefaced by S, but will be drawn from this volume.

[108] Provided it is not an instance that we recognize as one in which we need not or ought not consult conscience.

In a sense, this second way for a Butlerian agent to be vicious is just a species of the first, since failing to consult conscience is just a violation of one of the many obligations that conscience enforces over agents. Though we may only suspect (and thus not know) what conscience will dictate concerning the action we are considering, we already know (once we are acquainted with it) that conscience dictates that we consult it with regard to all our action.[109] When we know this and yet fail to act in accordance with it, we are simply being vicious in the first way described. And this is a way that we have already seen does not threaten the authority of conscience.

There is a third way to be vicious. It is also a case in which an agent does not consult conscience. But unlike the two ways considered above, it can neither be thought to be a case of failing to act on a known dictate of conscience, nor can it be thought to pose no threat to conscience's authority.

Notice that as I described the second type of viciousness, I was careful to qualify it to apply only to agents who are already acquainted with conscience. These agents are already aware of the faculty they have at their disposal and their obligation to avail themselves of it. But for agents who are not yet acquainted with either their faculty of conscience or their obligation to exercise this faculty, it is less than immediately obvious that such an obligation can exist for such agents. This possibility does indeed threaten the authority of conscience, if there are agents whom conscience cannot sensibly be thought to legitimately obligate.

Undoubtedly, agents to whom the obligation to reflect might not extend will be uncommon. In order to be agents, they will have to be endowed with the faculty of

[109] Though, as we have seen, it does not require that we consult it with regard to every instance of a type of action or immediately prior to our acting.

conscience; but they will have to be unaware of this faculty, never having used it or even noticed its presence in their psychological toolkit. It seems quite right to insist that such agents will be extremely rare, but there is no reason to suppose that there could not be such agents on Butler's account. He is already quite comfortable in admitting that there are brutes, agent-like creatures who act and survive solely on their own impulses, never appealing to conscience because they lack a conscience to which they may appeal. It is quite possible for brutes and agents to get by and survive without ever reflecting, so we may not exclude the possibility of these morally naïve agents[110] either in theory or in practice.

Once we admit that there could be naïve agents, we can see how they might pose a problem for Butler's account. The problem arises from the fact that the obligation to reflect is a moral obligation, but only the faculty of conscience is able to recognize moral obligations. Butler says as much in the first paragraph of his second sermon

> Now obligations of virtue shown, and motives to the practice of it enforced, from a review of the nature of man, are to be considered as an appeal to each particular person's heart and natural conscience, as the external senses are appealed to for the proof of things cognizable by them. (S.II.1)

The obligation to engage one's conscience is directed at, and can only be detected by, the faculty of conscience itself.

Here it seems that two of Butler's central claims might come apart. He intends obligations and responsibility to apply to agents merely by virtue of the fact that they possess the faculty of conscience. He elsewhere insists that we do not (and presumably should not) hold people responsible for things "which we consider as altogether out of [their] power." (DV.2) Naïve agents do indeed have the faculty required to be subject to

[110] For brevity, I'll hereafter call them "naïve agents".

obligations, but since they are entirely unacquainted with the faculty that they must engage in order to become aware of these obligations, it seems that meaningfully abiding by or shirking these obligations *is* altogether out of their power. Thus, we may not hold them responsible to these obligations, since such obligations do not extend to them.

In the passage where Butler notes that the obligation to reflect is addressed to the faculty of conscience he makes an analogy to the senses, particularly the eyes. Expanding this analogy will make vivid why obligations may not to extend to naïve agents. Imagine that there was an obligation to keep one's eyes open, but this obligation could only be expressed visually. We might imagine a world covered in placards, billboards, and neon signs urging us to open our eyes, so that we cannot help but become aware of this obligation once our eyes are open. But, if we have (for whatever reason) lived thus far with our eyes closed, it seems that we cannot be held responsible for failing to meet our obligation to open our eyes. We do not and could not know better, since it is only through our vision that we might become acquainted with this obligation. Waving the placards before us or pointing us toward the billboards will do no good until we are employing the faculty that allows us to notice these things. If we must be employing this faculty in order to see that we should be employing it, how can we be held responsible for neglecting an obligation we could only detect if we were using a faculty we do not even know we have? How can we be obliged to notice something we cannot see? In cases like these, the profound ignorance of a faculty of seems tantamount lacking that faculty altogether. A prisoner in a cell with a secret exit that he never finds is just as trapped as a prisoner whose cell has no exit to be found. Though their circumstances are metaphysically distinct, they are functionally equivalent; and so it seems we here ought to

hold naïve agents no more responsible than brutes, not relative to the sort of being they in fact *are*, but relative to the sort of being they are capable of *functioning as*.

So long as an agent is acquainted with her conscience she can be held accountable to the obligations that it points out to her: indeed, this is how she has such obligations. She will have seen the signs and billboards, as it were. But, if an agent is simply unaware that she has a conscience, it seems that the obligation to reflect cannot apply to her. This will mean that conscience cannot be universally authoritative, because there can be agents it cannot legitimately oblige. Not only will this be a problem for the consistency of Butler's own claims, it also threatens the categoricity requirement on morality's demands.

VI. Establishing Authority through Education

I have just argued that there is one way in which a Butlerian agent can be vicious, and though this sort of vice does not force us to conclude that the vicious agent is not an agent, we may have to concede that the agent cannot be held morally responsible for this vice. In this section I wish to preserve the claim that naïve agents cannot strictly be held responsible for failing their obligations, but that we may still have ways of bringing these agents under obligations. One of these ways even justifies treating naïve agents as if they were morally responsible for their actions.

I am unwilling to concede that we can legitimately consider naïve agents to actually be morally responsible for failing their obligations. Not only are naïve agents unaware of what their obligations are, they are unaware that there are even such things as

obligations.[111] Though they possess the capacity to reflect, so long as they are unaware of it, it is as inaccessible to them as it is to the brutes who lack it altogether; so, just like brutes these naïve agents have no such obligations. Naïve agents provide the limiting case to which Butler's counterfactual account of moral responsibility can extend.[112]

While I wish to retain this claim, I also wish to abandon an implication of the previous section concerning the analogy to vision, namely the idea that naïve agents could not come to be properly within the scope of obligations by becoming aware of their faculty of conscience. Again I can use the analogy to vision to make this point vivid. It is indeed useless to invest in brighter neon or to wave harder the placards before those with their eyes closed. Visual appeals to the (permanently or effectively) sightless are doomed. But the sightless have other faculties to which we may appeal; and these faculties might also have the power to open an agent's eyes.

I have granted that the *obligation* to open one's eyes can only be expressed visually to keep the analogy to conscience tight, but there might be other motives to urge opening one's eyes, just as other faculties might lead us to reflect. Given this constraint, I cannot simply *tell* an agent that she is obliged to open her eyes, but I might get her to exercise this faculty for other reasons. I might suggest that the films she so enjoys listening to are even better if she opens her eyes, or I could simply whisper in her ear that

---

[111] If Stephen Darwall's characterization of Butler in "Self-Deception, Autonomy, and Moral Constitution" as a constitutionalist is correct, it is more proper to say that there are no obligations for such agents, to avoid implying that obligations are metaphysically independent entities that conscience detects. Though I suspect Darwall is right on this score, I am reluctant to adopt this way of speaking for two reasons. First, I am still uncomfortable reconciling this view with Butler's habit of referring to conscience as a "moral faculty of perception". Second, speaking this way does serious violence to my analogy to vision, which is helpful to illustrate the points I wish to make here. Ultimately, I need not settle this issue here, since whether or not obligations are metaphysically independent of conscience or not, they will not apply to naïve agents.

[112] Perhaps what this actually indicates is that we should read Butler's talk about "capacity" functionally instead of as referring to a kind of metaphysical disposition.

she might enjoy exercising her eyelids, for the sheer physical pleasure of it. Once her eyes are open, she can see the obligation to open them herself, and the obligation holds over her now that she is aware of it. Simply because it is only her vision that allows her to see that there is an obligation to open her eyes does not require that there are not other means by which we can get her to open them.

It is through an insight like this that the moral education of even naïve Butlerian agents is possible. Fortunately for Butler's theory, this is precisely how much moral education actually takes place, by appeal to faculties that naïve agents such as children are already employing. We generally do not calmly ask four-year-olds to reflect on the virtues of reflection. Instead, we might appeal to their self-love and present them a task in which reflection is rewarded. We get them to begin reflecting on the grounds that reflection promises the reward we have attached to it, but once the child is reflecting and thus reflective, we may engage her according to this faculty. In this way, we are slowly getting them to function as the sort of creatures they are.

One way of provoking reflection is particularly suited to our purposes of salvaging moral responsibility. It gives us reason to treat naïve agents as if they were morally responsible up to a point, even if we would concede that they are in fact not.[113] Since naïve agents have no obligations, we hardly seem justified in wholeheartedly praising them or vigorously punishing them; but some sorts of moral praise and blame seem warranted. The warrant does not derive from the fact that that an obligation was genuinely met or shirked, but because the praising or blaming behavior might spark a naïve agent's conscience into action.

[113] My thanks to Steve Darwall for providing me with this example.

It is easiest to see how this might work when we consider blaming the ill deeds of naïve agents as wrong as the violation of an obligation, even though we know they are not. In blaming the naïve agent, we might insult her pride or threaten her self-conception. It is natural to assume that once a proud agent's justification for an act has been questioned, she will try to defend herself by justifying her actions. But to do this she will reach for the faculty by which we consider actions and justify them, conscience, and then will begin reflection. By holding naïve agents morally responsible even when they are not, we may get them to appeal to the faculty of which they were unaware. We can turn functional brutes into agents this way. Once they appeal to it, their obligations in light of it will become vivid to them, they can then legitimately be held responsible to these obligations, because now, for the first time, they genuinely *have* obligations.

Naïve agents may be distinguished from all other agents in that they are genuinely not morally responsible to respect obligations. Nevertheless, we still have reason to treat all agents as (more or less) morally responsible for obligations, because they either are morally responsible or are likely to become so when we treat them as such.

VI. Conclusion

Given these reflections, it is clear that Butler does provide a viable account of bad action. A being is an agent if she possesses the capacity to reflect and is vicious if she does not properly use or abide by this capacity. Although we are not justified in claiming that agents are morally responsible for their viciousness in all cases, we still have reason to treat agents as morally responsible in all cases.

This discussion began as an investigation into whether or not Butler suffered from the same difficulties as other theorists in his tradition. At least in terms of providing a viable account of bad action, he does not. He is able to avoid the pitfalls of Kant's and Korsgaard's views, by presenting standards of action and standards of morality that can coherently come apart. Thus, not all those in the autonomist tradition are unable to account for vicious agency. However, it does not appear that Butler's success in this area can be translated into considerations that can aid either Kant or Korsgaard.

We have seen how Butler might be pushed to concede that not all agents are morally responsible in all cases, because there can be some agents who, properly speaking, have no such obligations. This is a conclusion Korsgaard will surely wish to avoid, since she pitches her account against the most committed skeptic, and intends her obligations to extend to all agents, whether they are aware of, sensitive to, or willing to recognize these obligations. Her aspirations to meet the categoricity requirement militate against her accepting such a conclusion.

But beyond this relatively minor commitment to the scope of obligation, neither Kant nor Korsgaard can help themselves to the feature of the Butlerian account that allows for bad actions. Butler can do so because one is a Butlerian agent in virtue of possessing a capacity whether or not it is exercised. Conversely, Kant and Korsgaard appear to admit even a smaller class of beings into the community of agents, for agency only extends as far as the instances in which the capacity for agency is exercised. In order to align themselves with the Butlerian notion of agency, Kant would have to admit that the heteronomous actions of beings with the capacity to be autonomous were actually

free both neutrally *and* rationally (and therefore autonomous). He would have to make the two senses of "freedom" that Sidgwick isolated actually become one.

Korsgaard would have to admit that even beings who could but do not act on self-constituting principles are even then still agents. This would mean that, contrary to her argument to establish its impossibility, particularistic willing would count as a kind of willing and a kind of self-constitution. And since there are agents who can act on the basis of particularistic principles, they will not be subject to the demands of morality. Thus morality will not be just a matter of proper agency.

But these changes do not look promising. The suggested alterations to Kant's theory look to be impossible; those suggested to Korsgaard's look to make it another view altogether. Thus they are not changes that either theory could, in one way or the other, survive.

What more general lessons can be drawn from Butler's account and its comparison to Kant's and Korsgaard's is the final subject to which I now turn in my last chapter.

Chapter 5

Concluding Remarks:
Lessons and Horizons

I. Ethical Constitutivism

i) Constitutivism's Mistake

Constitutivism's mistake with respect to the fallibility constraint, at least, is grounding normative authority in precisely the same feature of an action that determines its kind membership. If one acts well, on the general constitutivist picture, then one has done precisely as one ought; but when one "acts" poorly, no obligation has been violated, since in "acting" badly a creature ceases to be the sort of thing subject to the obligations of agency, because it is no longer an agent. For Kant, this happens when one "acts" heteronomously (and therefore badly), but the act is to be attributed to something other than the agent. For Korsgaard, this happens when one "wills" particularistically (and therefore badly), and thus there is simply no agent to be found in the context of the action: it is indistinguishable from the interplay of psychological forces.

Of course, Korsgaard does not take herself to have proposed a moral theory on which there is no bad action and is thus in violation of the fallibility constraint that she is only too happy to bring to bear against other moral philosophers. She believes that she has salvaged the possibility of bad action by locating the source of normative force in a

demand that can be satisfied in matters of degree.[114] But, as I have argued at the conclusion of Chapter 3, even if Korsgaard had managed to ground normative authority in a demand that can be satisfied more or less well, this does not guarantee any normative force that will urge us to satisfy the demand any better than we already have.

It should be no surprise, then, that Bishop Butler does not manage to satisfy the fallibility constraint through the use of a constitutive standard that admits degrees of satisfaction. This is not entirely correct because Butler's view involves two distinct constitutive standards. The constitutive standard for what it is to be an agent is an all or nothing matter: either God gave you a conscience or He didn't. If you have one, you are an agent. If not, you are a brute (you're only a brute if you are animate, otherwise you could just be a rock).[115] The constitutive standard for what it is for an agent's actions to be *good* might well admit degrees of satisfaction, since consulting conscience is a deliberative process that extends through time and can be done more or less thoroughly.[116] But this capacity to admit degrees of satisfaction does nothing to help Butler's view accommodate the fallibility constraint; instead what does help is the fact that he is employing two distinct constitutive standards here. So long as they are separate, it will not matter if they are binary or continuous.

---

[114] It is on this point that I most squarely disagree with Korsgaard. Contrary to her own understanding of her view, I take it that what is constitutively normative for action on her account is an all or nothing matter, and what can be satisfied by degree need not be normative for action, and thus does not constitute it at all.

[115] As for what Butler should say about those who have a conscience that is unbeknownst to them is that they still *count* as agents, but might be excused from their obligations because they cannot yet *function* as agents. This is a particularly appealing answer because it seems to be just what we ought to think about young children. They are in our kind, just not yet up to our speed.

[116] Presumably, we should read Butler so that we can, at some point, stop considering every possible arcane implication of our action and simply notice that it is wrong to steal. Similarly, we might have consulted conscience and noticed that it is wrong to take things that do not belong to you, and then decided against such a taking, before waiting to notice further that this was an instance in which the taking was justified, perhaps even obligatory, and we thus came to the wrong conclusion even though we did consult our conscience. It seems that we can reflect more or less well, and so we should allow that following our conscience can be a matter of degree.

On Butler's account, the way to be a bad agent, or at least to have performed a bad action is to have been given a conscience by God (and thus satisfied the constitutive standard of agency), but to have failed to have consulted that conscience before acting (and thus failed to satisfy the constitutive standard of right action). It is because these constitutive standards are distinct that they can be satisfied independently and something can thus simultaneously count as both "wrong" and an "action."[117]

What comparing Butler's moral philosophy to Kant's and Korsgaard's has shown us is that it is a dangerous mistake to have the same constitutive standard determine both what something *is* and what it *should be*. (Even putting it this way suggests the way that nothing could go wrong on such an account.) Thus it appears that it is only the constitutivist strain in Kantian ethics that is generating difficulties against the fallibility constraint, because other views, like Butler's, which anticipates and bears a strong family resemblance to the Kantian tradition, do not have the same difficulties because they do not embrace the same constitutivist aspirations.

ii) Did the Butler Do It?

Bishop Butler's ethics and moral psychology do succeed at accommodating the fallibility constraint, but I hesitate to recommend here Butler's strategy as the one that all moral philosophy should adopt. This is because his view is still fairly far from accommodating all the intuitive constraints on a moral theory. I will pause here to note just two of them.

[117] I don't think Butler can give us something simultaneously counts as "right" and a "mere behavior." This will be because all the apparent instances of "right behavior" will have to have been preceded by some conscientious reflection, even if it is in the distant past. For Butler, brutes can help or hurt, but can do no right or wrong. Which, again, seems the right thing to think about animate creatures without a conscience.

The first is a difficulty with Butler's view that any constitutivist would likely point out, because it is a problem from which constitutivism does not suffer, and might even be thought of as being designed to solve. This is the difficulty in satisfying what I have called earlier the "categoricity requirement." It is the idea that moral obligations apply equally to all similarly situated moral agents, and that none can legitimately avoid the authority of morality. Both Korsgaard and Velleman are quick to point out that constitutivist theories provide excellent replies to skepticism, because it is obvious *why* you should ξ in order to ϕ, because ξ-ing is just *what it is to* ϕ.

It seems fairly clear at least to me that what is appealing about this view is also what attracts people to ethical internalism. In both cases, to have a genuine instance of something, you are guaranteed to also have the feature you are trying to capture. So, for Korsgaard's constitutivism, every time you have an agent, you automatically have a self-constitutor committed (though she may not realize it) to unifying herself according to the Categorical Imperative. For internalism about moral reasons, every time you have someone genuinely judging that something ought to be done, you will also have an agent who is *motivated* to do what they judge ought to be done. Skepticism about why one should self-constitute or be motivated by one's judgments about what is right are just *impossible* within the framework of these sorts of accounts. Skeptics reveal their howlingly bad understanding of agency or moral motivation when they raise such doubts. Korsgaard is quite right to locate herself in a tradition extending down from Plato, because both her constitutivism and this sort of judgment internalism look to just be descendants of the Platonic doctrine that "to know the good is to do the good."[118]

[118] See Plato, 1997. *Protagoras* 358c and *Meno* 77c and 78b. All references to Plato are to this collection and cite Stephanus pages.

Butler is able to accommodate the fallibility constraint precisely because he is willing to admit that there are agents who know the good, yet who fail to do it. The kinds of viciousness that are paradigm satisfactions of the fallibility constraint on Butler's account are when an agent either consults her conscience and then fails to follow its dictates, or fails to consult it, fearing what it might recommend. These cases can be a source of joy if one is seeking ways to satisfy the fallibility constraint, but they are vexing puzzles for those trying to solve the mystery of moral motivation. In short, the way Butler is able to give a coherent reply to the question "How can I be bad?" is also the way he will be flummoxed by the question "How can I be good?"

Once a Butlerian agent has consulted his conscience and found out what to do, it does not seem that there is anything more we can offer him in terms of motivation. If seeing and clearly understanding his very obligation to do something is not enough to provide him moral motivation, we may be up a river in terms of solving the mystery of moral motivation.[119]

As far as the three cases I have examined suggest, satisfying the fallibility constraint might only come at the cost of the inability to meet the categoricity requirement, and vice-versa. I have a hunch that this tension runs much deeper than in just the specific cases I have examined here, but I must leave this suspicion to later and more expansive investigation.[120]

That is the substance of my first reason to be reluctant to widely recommend Butler's solution as a way of generally meeting the fallibility constraint. My second

[119] I am here being careful to say "moral motivation," because we could, of course, offer bribes or threats to get the agent to do what he should, but these incentives are equally available to brutes and thus do not constitute the distinctively *moral* motivation we are trying here to capture.

[120] My thanks to Joshua Brown and Frank's Restaurant for providing the habitat and opportunity to develop this hunch.

reason has to do with worries about teleology, and looks to be endemic to both Butler's and Korsgaard's accounts.

In an unpublished but fairly widely circulated manuscript, Gideon Rosen has raised doubts about the kind of normativity one can draw out of constitutive standards in the first place. To put his suggestion into a slogan, it seems that constitutive standards can only generate normativity in the sense of things being *correct* or *incorrect*, they cannot extend to the different, and further, sort of normativity that dictates what *should* or *should not* be done. Correctness conditions, while normative, do not settle questions of what ought to be done. Paraphrasing Rosen's own example, the correctness conditions for a playing of Mozart's C-major piano sonata are settled by its score. So, if you play C# where a B is indicated by the score, you have played the sonata *incorrectly*.[121] The score is (let's stipulate) the final arbiter of what constitutes correct or incorrect playings of Mozart's sonata, but the score cannot settle the different, and further, question of *whether you should play the sonata correctly or not*. Say I wish to make a musical joke, or conceal my piano skill from someone who would be made jealous by it. Now it looks as if I have ample reason to play the sonata *incorrectly*. As Rosen observes, these reasons of mine will not make my C# the *correct* note in the sonata, but it might nevertheless be a note I *ought* to have played, given my circumstances and reasons.

It is considerations like this that lead me to conclude that when Alyssa sets out to defy the rule of the avant-garde art example, she is engaging in a kind of following. The score of Mozart's sonata tells me which notes are correct, but I can then follow that information in order to play the notes correctly *or* incorrectly. In both cases the score functions as my guide or map, but only I get to decide where I'm actually going to go.

[121] See Rosen, Manuscript, pp. 10-11.

The only wrinkle in Alyssa's case is that she is setting out to make it to a place that doesn't exist (i.e. a submission that will not satisfy the rule). This gives us the further result about unbreakable rules that even though they cannot fail to be satisfied, they are not therefore unavoidable sources to settle the question of what we *ought* to do.[122]

Rosen's criticisms are just aimed at constitutive standards, but I believe they can be expanded to any sort of teleological standard, even those like the one Butler employs. Korsgaard establishes the normative authority of self-constitution by appealing to the brute fact that we *just are* creatures that self-constitute (for we are agents) and so this is why we *ought* to constitute ourselves: it is impossible to do otherwise. Butler establishes the normative authority of conscience in a different way. True to his historical context (and his being a Bishop) he appeals not to brute facts, but to God. The reason we should obey our conscience is because, when we consider our constitution we can observe that God *designed us* to consult and obey it. As he says in the Preface,

> [I]n reality the very constitution of our nature requires that we bring our whole conduct before this superior faculty, wait its determination, enforce upon ourselves its authority, and make it the business of our lives, as it is absolutely the whole business of a moral agent to conform ourselves to it. (S.Preface.25)

Since we know that God designed us deliberately, we can infer from the pride of place He gave conscience in our constitution that we ought to be using it to determine the conduct of our lives.

There is the obvious worry that this solution of Butler's depends on a proof not only for the existence of God, but also for the truth of something like the teleological

[122] I take it that this point undermines the strength of Korsgaard's appeal to the "necessity of action." Even if acting is unavoidable, we can still decide whether to struggle against this fact or go along with it, as Alyssa does with the art assignment. Even if we are doomed to act, this does not settle which act we ought to do.

argument for His existence. Beyond this is the claim that is subject to just the same sort of worry as Rosen's complaint against the idea that all normativity can be derived from constitutive standards.

This is because both Korsgaard's claims and Butler's claims are fundamentally *teleological* claims. Korsgaard thinks it is *just true* that we are beings who self-constitute and so we should do so. Butler, instead, thinks that we are creatures who use our conscience to guide us because God has designed us so to be, but we nevertheless *should* use our conscience because we were so designed.

Any reader of Hume should be worried at this point, because Korsgaard and Butler have both moved from an "is" to an "ought." And it seems that this is what all teleological claims, of which constitutivist claims are a subset, do.

Some are tempted to read the history of Western moral philosophy as a crisis and scramble to hold up the moral claims that used to be supported by God and the teleological universe He designed, but are now threatened to fall in the naturalistic picture that lacks a God and teleology to hold them up. But I think this is a mistake, perhaps not of what historical figures have understood themselves to be doing, but certainly as a description of what is the case. God is not able to decree things into being right, and neither is teleology capable of grounding oughts. If this is all that was holding morality up in the first place, it has been in free fall all along.

We can use Rosen's insight to see that neither God nor Aristotle was capable of establishing the sort of "oughts" and "shoulds" we seek in moral philosophy. At most, God can settle what He thinks is correct, and could create features and things with a purpose in mind, but this could, at best, generate something analogous to correctness

conditions for these objects, it would not settle the question of what I ought to do. Simply because it is an *incorrect* way to fire the gun to squeeze its barrel instead of its trigger, this does not mean that this isn't precisely what I should do. Similarly, simply because the Barry Manilow album was made for playing in my stereo, this does not mean that I should listen to it.

Because all teleological claims look, at best, to be able to generate the normativity of correctness conditions and not the normativity of "oughts" and "shoulds," it seems that we will have reason to shy away from any such attempt to ground morality in essentially teleological claims, be they constitutivist or otherwise.

II. Other Normative Horizons

As I have mentioned in passing earlier in this dissertation, the fallibility constraint has cognates in subfields of philosophy other than ethics. My discussion here has focused on ethics and practical reason, but the possibility of failure seems tied up with normativity itself in such a way that we should expect it to play a role in all our normative domains. As Nicholas Rescher puts it, "The three prime spheres of human concern are belief, behavior, and evaluation, which correlate with matters of fact, action, and value. And one can manage to err in all three settings."[123]

As this dissertation will indicate, I suspect that this is not because of some deep feature about the nature of rules, or about the necessary threat of alternate possibilities. Instead, I believe that we will have reason to insist on a fallibility constraint whenever we are independently committed to failed instances of a kind.

[123] Rescher, 2007, p. 1.

Obviously, establishing this claim is beyond the scope of my current project, but I would like to close with one promising piece of evidence from thinking about the semantics of mental content in the philosophy of mind. Semantic considerations are what fundamentally drove Wittgenstein, Kripke, and Rosen, so it is fitting that I should return to them to close this dissertation.

In *Psychosemantics*, [124] Jerry Fodor engages a well-known difficulty with the causal theory of content. Specifically, it seems that any causal theory of how concepts get their content will have a terrible time accounting for instances of misrepresentation. There are multiple discussions and formulations of this problem, but Jerry Fodor's treatment of it in *Psychosemantics* is perhaps most useful for present purposes thanks both to his concise statement of the problem, and his fascinating proposed solution to it.

Roughly, a theory of content is a causal one if it claims that the content of a concept is determined by whatever causes that concept to be "tokened" (i.e. provoked) in an agent's mind. This is enough to generate the problem with accounting for error, because it means that whatever provokes a concept in my mind is what that concept means. So, it turns out that I can't walk into a dark barnyard, see a cow, and *mistakenly* think "horse," because my concept "horse" now means "horses or that cow" because that cow just managed to provoke my concept "horse". What should have been a straightforward case of misidentification was actually a more obscure case of concept revision, and it looks like all potential mistakes will have to be just such concept revision, according to a causal theory.

Fodor himself considers what he calls the "Crude Causal Theory (CCT)" which "says, in effect, that a symbol expresses a property if it's nomologically necessary that *all*

[124] Fodor, 1987. Hereafter referred to parenthetically in the text as P with page number(s).

and *only* instances of the property cause tokenings of the symbol." (P 100) As Fodor puts it, this leads to "[a]n embarrassment: It seems that, according to CCT, there can be no such thing as *misrepresentation*." This is because, granted that the property A causes the symbol 'A', then

> the tokens of the symbol denote A's (since tokens denote their causes) and they represent them *as* A's (since symbols express the property whose instantiations cause them to be tokened). … So it seems that the condition for an 'A'-token meaning *A* is identical to the condition for such a token being true. (P 101)

So far so good, until we get to a case of misrepresentation, say when something that is not an A, say, a B, causes a tokening of 'A'. Then,

> it follows that the causal dependence of 'A's upon A's is imperfect; A's are sufficient for the causation of 'A's, *but so too are B's*. If, however, symbols express the properties whose instantiations reliably cause them, it looks as though what 'A' must express is not the property of *being A* (or the property of *being B*) but rather the *disjunctive property of being (A v B)*. (P 101)

Fodor calls this "the disjunction problem." (P 102)

Fodor then devotes his efforts to resolving the disjunction problem, but it is worth a moment to pause and consider the idiosyncrasies of this view and as well as its dialectical location.

First, it is interesting to notice the way this version of a difficulty in accounting for error works differently from others we have considered. The difficulty with Humean practical reason, according to Korsgaard, was that there simply were no candidates for erroneous actions: whatever you did was just what you had most reason to do, according to this theory. One difficulty with Korsgaard's constitutive account, in my assessment, is that it refuses to recognize candidates for erroneous actions as actions: if your behavior is not "universal" it simply doesn't count as an action and thus cannot be an erroneous

action. The disjunction problem, conversely, eliminates its candidates for erroneous representations by assimilating them into *successes*: if a B causes me to think 'A', this is no case of my mistaking a B for an A, it just instead makes it the case that 'A' turns out to mean something different that I originally thought, namely (A v B). These differences may indicate that there are at least three distinct ways to fail to account for error.

Secondly, it is no surprise that Fodor devotes all of his efforts to framing and replying to the disjunction problem, and devotes no space to explaining *why* a theory of mental content must be able to account for misrepresentations. This is because, in this context at least, the reason is so obvious that it does not bear mentioning. Our theory of mental content must be able to account for misrepresentations because we already know that *misrepresentations occur*. Even if we've never personally misidentified a cow as a horse, such errors are ubiquitous and are obviously a piece of data that any credible theory of mental content will have to be able to account for, if it is to adequately describe the phenomenon of mental content.[125]

This is just what I would have us expect given the structure of my justification of the fallibility constraint. We have reason to insist on the possibility of a particular kind of failed moral agency, I suggest, because we are acquainted with such a thing in the

[125] Though it does not warrant engaging in the text at the moment, I cannot help but mention a feature of Fodor's solution to the disjunction problem. In order to ground an asymmetry between accurate representations and misrepresentations (or "wild tokenings"), he claims that misrepresentations causally depend on the causal route for accurate representations, but not vice versa. This notion, Fodor derives from "an old observation ––as old as Plato, I suppose–– that falsehoods are *ontologically dependent* on truths in a way that truths are not ontologically dependent on falsehoods. … you can only have false beliefs about what you can have true beliefs about (whereas you can have true beliefs about anything that you can have beliefs about at all.)" (P 107) This is especially interesting for my purposes because here Fodor is claiming that, in the normative context of truth and falsity, the normativity of true beliefs is entirely independent of any failures, but instead *failures themselves* (i.e. false beliefs) are ontologically dependent on the normativity of truth. If Fodor is correct, we now have a reversed case of the error constraint, that there can be no failures if there is no normativity. (Perhaps what is most surprising about this is Fodor's insistence that you can have true beliefs without any dependence on false ones. The fact that there's no failure without normativity is not news.)

world. Just as in Fodor's discussion, there is an onus on any theory of content to be able to account for a misrepresentation, because we are acquainted with misrepresentations in the world. We ignore these facts at our peril, and if we propose normative theories on which no such failures can occur, it will be those theories that have made the mistake.

Bibliography

Anscombe, G.E.M. 2000. *Intention* (Cambridge, Mass.: Harvard University Press).

Aristotle, 1941. *Nicomachean Ethics* translated by W. D. Ross in *The Basic Works of Aristotle*, edited by Richard McKeon (New York: Random House).

Arpaly, Nomy. 2003. *Unprincipled Virtue* (Oxford: Oxford University Press).

Broome, John. 1993. "Can a Humean be moderate?" pp. 51-73 in *Value, Welfare, and Morality* eds. R.G. Frey and Christopher W. Morris (Cambridge: Cambridge University Press).

Butler, Joseph. 1873. *Sermons*, (New York: Robert Carter & Brothers).

Butler, Joseph. 1983. *Five Sermons Preached at the Rolls Chapel and A Dissertation Upon the Nature of Virtue*, (ed. Stephen L. Darwall), (Indianapolis, Ind.: Hackett).

Darwall, Stephen L. 1988. Self-Deception, Autonomy, and Moral Constitution" pp 407-430 in *Perspectives on Self-Deception*, McLaughlin, Brian P. and Rorty, Amélie Oksenberg (eds.), (Berkeley: University of California Press).

Fodor, Jerry, 1987. *Psychosemantics* (Cambridge, Massachusetts: The MIT Press).

Frankfurt, Harry G. 1988a. "Alternate possibilities and moral responsibility," pp. 1-10 in *The importance of what we care about* (Cambridge: Cambridge University Press, 1988).

Frankfurt, Harry G. 1988b. "Rationality and the unthinkable," pp. 177-190 in *The importance of what we care about* (Cambridge: Cambridge University Press, 1988).

Herman, Barbara. 1993. *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press).

Hobbes, Thomas. 1994. *Leviathan* ed. Edwin Curley (Indianapolis: Hackett Publishing Company, Inc.).

Hume, David. 1978. *A Treatise of Human Nature* eds. L.A. Selby-Bigge and P.H. Nidditch. (Oxford: Oxford University Press).

Kant, Immanuel. 1996. *Practical Philosophy* ed. Mary J. Gregor (Cambridge, Cambridge University Press).

Kant, Immanuel. 1997. *Lectures on Ethics* ed. Peter Heath and J.B. Schneewind; trans. Peter Heath. (Cambridge: Cambridge University Press).

Korsgaard, Christine. 1996. *The Sources of Normativity*, (Cambridge: Cambridge University Press).

Korsgaard, Christine. 1997. "The Normativity of Instrumental Reason," pp. 215-254 in *Ethics and Practical Reason* eds. Garrett Cullity and Berys Gaut (Oxford: Oxford University Press).

Korsgaard, Christine. 1999. "Self-Constitution in the Ethics of Plato and Kant," *The Journal of Ethics* 3: 1-29.

Korsgaard, Christine. 2002. *Self-Constitution: Action, Identity, and Integrity* Manuscript of Locke Lectures previously available at her website, http://www.people.fas.harvard.edu/~korsgaar/#Publications .

Korsgaard, Christine. 2009. *Self-Constitution: Action, Identity, and Integrity* (Oxford: Oxford University Press).

Kosch, Michelle. 2006. *Freedom and Reason in Kant, Schelling, and Kierkegaard*, (Oxford: Oxford University Press).

Kripke, Saul A. 1982. *Wittgenstein on Rules and Private Language*, (Cambridge, MA: Harvard University Press).

Lavin, Douglas. 2004. "Practical Reason and the Possibility of Error," *Ethics* 114 (April): 424-457.

Millgram, Elijah. 2005. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*, (Cambridge: Cambridge University Press).

Plantinga, Alvin. 2007 "The Free Will Defense," pp. 315-340 in *Philosophy of Religion: Selected Readings*, eds. Michael Peterson, William Hasker, Bruce Reichenbach, and David Basinger (Oxford, Oxford University Press.) Selected from Plantinga's *God, Freedom, and Evil*, (Grand Rapids, Michigan, Wm. B. Eerdmans Publishing Co., 1977).

Plato, 1997. *Complete Works* ed. John M. Cooper (Indianapolis: Hackett Publishing Company).

Rachels, James. 1975. "Active and Passive Euthanasia," *The New England Journal of Medicine* Vol. 292, January 9, pp. 78-80.

Railton, Peter. 1997. "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action," pp. 53-79 in *Ethics and Practical Reason* eds. Garrett Cullity and Berys Gaut (Oxford: Oxford University Press).

Railton, Peter. 2006. "Normative Guidance," pp. 3-33 in *Oxford Studies in Metaethics: Volume 1*, ed. Russ Shafer-Landau (Oxford: Oxford University Press).

Rescher, Nicholas. 2007. *Error* (Pittsburgh: University of Pittsburgh Press).

Rosen, Gideon. Manuscript. "Normativity, Meaning, and All That." (Unpublished).

Ross, W.D. 2002. *The Right and the Good* ed. Philip Stratton Lake (Oxford: Oxford University Press).

Sidgwick, Henry. 1888. "The Kantian Conception of Free Will." *Mind* Volume XIII, Number 51: 405-412. Reprinted in his *The Methods of Ethics* (Indianapolis, Ind.: Hackett, 1981) pp. 511- 516.

Smith, Michael. 1994. *The Moral Problem* (Oxford: Blackwell Publishers).

Strawson, P.F. 1974. "Freedom and Resentment," pp. 1-25 in Strawson's *Freedom and Resentment and other essays* (London: Methuen & Co.).

Velleman, J. David, 1992. "The Guise of the Good." *Noûs* 26: 3-26.

Velleman, J. David, 2000. "The Possibility of Practical Reason," pp. 170-199 in his *The Possibility of Practical Reason* (Oxford: Oxford University Press).

Velleman, J. David, 2006. "So It Goes," The Amherst Lecture in Philosophy 1 (2006): 1–23. <http:// www.amherstlecture.org/velleman2006/> .

Watson, Gary. 2002. "Volitional Necessities," pp. 129-159 in *Contours of Agency* eds. Sarah Buss and Lee Overton (Cambridge, Mass.: The MIT Press).

Widerker, David and McKenna, Michael (eds.). 2006. *Moral Responsibility and Alternative Possibility: Essays on the Importance of Alternative Possibilities* (Aldershot and Burlington: Ashgate Publishing).

Wittgenstein, Ludwig. 1998. *Philosophical Investigations*, translated by G. E. M. Anscombe, Third Edition. (Oxford: Blackwell Publishers).

Wolf, Susan, 1980. "Asymmetrical Freedom," *Journal of Philosophy* 77 (March): 151-166.