Acceptance, Belief, and Partiality:

Topics in Doxastic Control, the Ethics of Belief, and the Moral Psychology of Relationships

by

Laura K. Soter

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy and Psychology)
in The University of Michigan
2022

Doctoral Committee:

Professor Ethan Kross, Co-Chair
Professor Peter Railton, Co-Chair
Professor Susan Gelman
Professor Renée Jorgensen
Professor Chandra Sripada

Laura K. Soter

lksoter@umich.edu

ORCID iD: 0000-0002-3653-8211

For my parents, who somehow let their only child become a philosopher.

ACKNOWLEDGEMENTS

Over the past few years, as I've watched a steady stream of Facebook friends receive their PhDs and post about their dissertations, I've picked up the habit of reading acknowledgments. This habit has steadily extended beyond friends' dissertations into books and other popular works.

Academics, and perhaps especially philosophers, are often trained to focus only on the ideas themselves as presented, and not to pay much attention to the individual who created those ideas. But acknowledgments reveal that there is a full person behind the ideas, and then many more people behind the person behind the ideas. Theses glimpses into other authors' lives and stories serve, to me, as a helpful reminder that ideas do not pop into the world fully formed, that an author is a person and not just a vehicle for ideas, and perhaps most importantly, that no one creates these projects alone.

So I find reading others' acknowledgments heartening—but also intimidating, because it has always made me feel like I would never be able to do justice to thanking the people in my own life. That worry has not abated, but the time has somehow come for me to attempt to thank the enormous number of people who have helped and supported me along the way. So I will do my best, with apologies for my inevitable inability to capture my gratitude with sufficient force and eloquence in these pages. (And also with apologies for the excessive length of these acknowledgments—I am lucky to have so many people who deserve these thanks!)

Before anyone else, I must thank my parents. They are the people to whom I owe everything, who have loved and supported me at every turn. Given my trajectory from an extremely moody teenager to starting, of all things, a PhD in philosophy, there would have been plenty of opportunities for that support to reasonably waver. But it never has. I don't have the space to list all the things I am grateful to my parents for here, so instead I will highlight just one that stands out: I have always felt that they genuinely trusted me to make my own life decisions, and that even when they didn't fully understand them (like studying philosophy), they really believe that I can succeed at whatever it is that I set my mind to. So, Mom and Dad—thank you for everything. I love you both so much.

Next, my wonderful committee. I have been so lucky to have the true dream team of advisors in both philosophy and psychology; I experience recurring disbelief at the people I get to work with, and I am so grateful to all of you for being my mentors over the years. Peter Railton and Chandra

Sripada have both served as the best possible role models for how to combine philosophical rigor with serious empirical engagement. They both helped me shape my ideas from pages of nebulous ramblings into a real-life dissertation; I am so grateful for the way they have encouraged me, guided me, and pushed me over the past few years. Peter has been a kind, thoughtful, and generous mentor; talking with him always leaves me with about ten pages of notes, a head full of ideas, and a renewed excitement for philosophy. He has helped encouraged me to be a big-picture thinker; I am constantly inspired by his ability to weave together ideas from such a broad range of disciplines and domains of philosophy. Chandra's no-nonsense advising has helped me gain both focus and confidence; I have always been grateful for his frank assessment of my ideas and straightforward advice. Like any grad student, I've dealt with my share of imposter syndrome and under-confidence, especially as I transitioned from writing course papers to trying to develop a dissertation. I learned early on that Chandra would tell me outright when he didn't like my ideas—which caused me to genuinely trust him when he said he *did* like something, and gave me the motivation to pursue it. It is thanks to Chandra that I have not just a couple of papers about acceptance, but something resembling a Research Project going forward; he has helped me learn to think more systematically and strategically about my work. Chandra also deserves a special thanks for talking me into a last-minute job market push this fall, and for sticking with me throughout the year as I navigated the process; his confidence in me (even when it felt totally undeserved) has helped me develop a lot more confidence in myself and my ideas over the past year.

I owe a special debt to Ethan Kross and Susan Gelman for agreeing to take a chance on me. They both welcomed me into their labs as an eager first-year philosophy student, and agreed to advise me when I put together the ridiculous plan of pursuing a joint degree. Ethan has taught me so much about how to think like a scientist: how to be practical and precise in my research questions, how to translate big ideas into concrete steps, and how to think through and justify the impact of my empirical work. Working with him also taught me a lot about communicating across disciplinary boundaries with psychologists; working with someone whose background is not in philosophy has been invaluable for my ability to navigate both disciplines. Nothing is more fun than being excited about a new project or dataset during a meeting with Ethan; I am grateful for his enthusiasm for research, and his optimism and positivity as an advisor. Susan, as anyone who has worked with her knows, is a truly special mentor. She is, of course, a rigorous and brilliant scientist, and she is also a caring and committed advisor. There is no one who is so absurdly busy and simultaneously so generous with her time and brainpower. I cannot count the number of hours I have spent meandering my way through pages of

half-formed research ideas with her; she has always been willing to help sift through the mess to try to find the kernel worth pursuing. She manages to strike the perfect balance of free-flowing exploration and practical guidance. At heart, she really is a philosopher: someone who is excited about the world of ideas, and I have been so lucky to learn from her.

Finally, special thanks to Renée Jorgensen for joining my committee at the very last minute and making it possible for me to defend this summer. Though the proximal cause of her joining my committee may have been logistics, I was thrilled at her addition; I had really wanted to work with her but thought that because I was wrapping up my degree, it didn't make sense to add her. In the brief time we have worked together, I have found her perspective incredibly helpful. I have been (semi?) joking that because I'm finishing up my time at Michigan earlier than anticipated, I've made all my advisors swear a blood oath that they'll keep working with me once I start my post doc. But really, this is because I am genuinely sad to be leaving this incredibly advising team behind. I cannot thank you all enough for your mentorship, support, and your willingness to work with me through the chaos of my joint degree.

I owe much to many of the other faculty and staff at Michigan as well. I am so grateful Maegan Fairchild for being the All-Around Best. Maegan has helped me in pretty much every conceivable way over the past year: she has helped me (re)structure papers and organize my thoughts in ways I struggled to do on my own; offered invaluable feedback on my philosophical work; been an enthusiastic sounding board for budding and chaotic ideas; worked with me on Philosophy with Kids; given feedback on syllabus and course design as I prepared to solo teach my first class; and offered incredibly helpful advice and emotional support as I navigated an unexpected job market push (sorry for all the frantic emails, Maegan!). Sarah Buss has consistently asked some of the hardest and most probing questions of anyone I've talked to about my work; thanks to her for urging me to think about depths of my projects I would not have arrived at on my own. Thanks to Brian Weatherson for his stint on my committee and all the helpful feedback and guidance he provided in that role. Laura Ruestche offered invaluable guidance and support in her role as Director of Grad Studies when I was early in the process of navigating my joint degree; I could not have done so nearly as successfully without her help. Dave Dunning taught approximately half of the courses I took in the psychology department; he has profoundly shaped how I think about social psychology, and I am grateful to him for welcoming me into his classes and the department. I am also grateful to all the other professors whose classes I took: Eric Swanson, Sarah Moss, David Manley, Victor Caston, Scott Hershovitz, Guillermo del Pinal, Liz Anderson, Rich Gonzalez, Felix Warneken, and Andrei Boutyline. A special thanks to Andrei for

taking a chance on me and inviting me to co-author with him on what we thought would be a quick paper, but which turned into a three-year undertaking. I learned so much through working with him, and had a great time doing so. Our paper also serves as an important success in my mission to make my CV as confusing as possible. Fan Yang has been another recent collaborator and mentor with whom I feel lucky to work. Additionally, thanks to all the faculty who served as teaching mentors: David Manley, Jim Joyce, Eric Swanson, and especially Mara Bollard. Mara, in particular, went out of her way to share her extensive teaching knowledge with me—and also spent a lot of time being a sounding board and source of advice during a stressful term—and I am very grateful I got the chance to work with her. Thanks to Jude Beck, Kelly Campbell, and Shelley Anzalone, who have been both friendly faces and helpful resources over the years. And finally, my endless thanks to Carson Maynard, the best grad coordinator anyone could ever wish for. I have put Carson through the wringer navigating the bureaucracy of my joint degree, and yet he has always been one of the most kindhearted and cheerful people I've ever had the pleasure of knowing. The philosophy department is so lucky to have him, and he has been a real bright spot in my time at Michigan.

I would never have made it to Michigan in the first place without the guidance of my mentors at Carleton College. The Carleton Philosophy Department is a special place. Carleton was where I first discovered and fell in love with philosophy, even if I stumbled into it somewhat by accident. In my first semester, I took Jason Decker's "Science, Faith, and Rationality" class. By the end, I couldn't decide if I'd loved it or hated it (the ideas were so interesting, but the readings were so hard and I was always so confused!!), so I took a second class with Daniel Groll—and after that I was hooked. Over the course of my undergraduate career, Jason and Daniel became formative mentors, helping me develop my philosophical skills—but also teaching me to be confident in myself and my ideas (and to stop panic-emailing them before every paper was due). I am so grateful that they put up with me as an over-enthusiastic, neurotic undergraduate who constantly bugged them in office hours. They made philosophy challenging, fun, and something I could see myself wanting to do. I would not be where I am without them. Doug Marshall also took me under his wing when he joined Carleton's department my sophomore year. I'm pretty sure that I took more classes with Doug than any other student in his first couple of years at Carleton—starting with being the terrified only sophomore in his Early Modern class full of senior majors. Doug's teaching and courses challenged me throughout undergrad and undoubtedly helped me grow, and I am grateful to him for all the time and energy he put into helping me learn over the years. Anna Moltchanova and Sarah Jansen were also integral parts of my Carleton

philosophy experience, and I am grateful to them for their teaching and conversation (as well as for letting me sleep on the department couch at all hours).

I had equally wonderful mentors in Cognitive Science and Psychology at Carleton. Kathie Galotti is the fierce leader of Carleton CogSci, and taught me so much of what I know about the field. She is the kind of teacher and advisor who has very high expectations of her students because she believes they can live up to those expectations, and I grew immensely as a student working with her. She is the person who first introduced me to the field of moral psychology—one of the first domains that made me really think, "*that's* what I want to do!" I was also lucky to work in Adam Putnam's lab studying collective memory. Adam taught me a lot about what it's actually like to do social psychology, rather than just read about it, and my experiences working with him were a large contributor in my interest in pursuing psychology in graduate school. Another testament to the unique intellectual environment and mentorship at Carleton can be seen in the number of people I overlapped with there who are also pursuing PhDs in philosophy or psychology. Among those I have seen around the academic world in recent years include Emily Tilton, Camila Hernandez Flowerman, Anna Smith, Soren Schlassa, Valerie Umscheid, Morgan Ross, Elliot Schwartz, and Caroline von Klemper. I'm so happy our paths crossed in undergrad, and I hope they continue to in the future.

I also owe a special thanks to Tony Chemero. The summer after my junior year, Tony gave me my first research assistant job in his lab at Cincinnati. (While there, I was lucky to work with his student Patric Nordbeck, who was so kind and tolerant of my endless questions.) Working in that lab was a transformative experience on a number of levels. First, it was the summer I began to think that maybe I really wanted to go to graduate school for philosophy, and Tony was a generous source of advice as I began to navigate the graduate admissions process. Second, Tony's lab studies radical embodied cognitive science and ecological psychology—a very different tradition than the world of classic social and cognitive psychology I'd been previously exposed to. Working in that lab was a dramatic demonstration of how philosophical commitments can profoundly shape how empirical science is approached, and though I may have slunk back to the world of beliefs, representations, and moral psychology, that perspective has stuck with me. Tony also gave me my first model of someone who was doing genuinely interdisciplinary work in philosophy and psychology: someone with formal training and active research programs in both disciplines. This planted the seeds for the joint program I would eventually seek at Michigan. I am extremely grateful to him for everything he has done for me since that first summer working in his lab.

I would not have made it through graduate school in one piece without the friendship and guidance of so many other graduate students in both of my departments. In philosophy, I was lucky to start my PhD with the world's best cohort: Angela Sun, Sumeet Patwardhan, Ariana Peruzzi, Katie Wong, Rebecca Harrison, and Cameron McCulloch. My cohort (the Philamily) has played a central role in my time at Michigan; they have been friends and sources of support, and graduate school would have been lonelier and less bright without them. Angela has been one of my main sources of laughter and solidarity at Michigan. Some of my favorite memories of grad school involve staying in Tanner until 3am with her writing papers, drinking both ciders and bottled coffee drinks, and of course crafting our manifesto, the Sun-Soter thesis. Angela is a confidant and a hype-woman; she is a light and I am so grateful for her friendship. Ariana is one of the people I most clearly remember from when I first visited Michigan. She is open and kind and genuine; I've learned a lot from her, and she's one of my favorite people to get distracted in the lounge with. Sumeet is simultaneously one of the most wise and competent, and also most silly people I know. He is an impeccably careful and well-ordered thinker who has taught me to organize my thoughts. He blew me away with his leadership of GEO through the early Covid days and the grad student strike. And he is also just a genuinely good friend; he's fun and playful and goofy and always down for a game night. Katie is maybe the person I talked the most philosophy with in my earlier years; I am grateful to her for letting me bounce ideas off her in Tanner. Rebecca always impressed me with the extent to which she knew her identity as a philosopher; she was a role model for confidence in her own ideas in my earlier years of grad school. And Cameron is just a fantastic person to get lost in conversation with; there is truly never a dull moment with him, and I really think he knows everything about philosophy.

Many other Michigan grad students have shaped my time as well. Adam Waggoner has been a committed co-organizer and a kind and generous friend. I have relied on him frequently when I overcommitted myself, and I'm grateful to him for everything. One instance of his generosity was bringing Maria Waggoner out to Michigan so that the rest of us could enjoy her company as well—I'm so glad I got the chance to know her over the years (and to meet Athan!). I suspect that making very close new friends in one's final year of grad school is somewhat rare; for that experience I'm grateful to Gabrielle Kerbel, who is an amazing friend and my favorite conference buddy. I never fail to get lost in conversation with Gabrielle for hours, about everything from *Selling Sunset* to difficult life questions to quantum physics; and though this feature of our friendship is excellent for my soul, it is terrible for my sleep. Gillian Gray is the best rock-climbing buddy I could hope for, and I am extremely grateful to her for never dropping me at the climbing gym and also for not judging my fear of heights.

Eduardo Martinez has been someone I looked up to since my first year of grad school. He has always been generous with his time and advice, from helping me prep for my first conference presentation in my very first year, to fielding endless job market questions in this final one. Josh Hunt has helped me create and sustain Philosophy with Kids, and has over the years given me endless helpful advice regarding navigating the profession, interdisciplinary work, and my taxes. Malte Hendrickx is both a great conference buddy and has been a cheerleader for me in a year when I frequently felt overwhelmed and inadequate, and I have been so grateful for his support and kind words. Emma Hardy has been a reliable friend and co-organizer, and I have always respected her ability to speak her mind with confidence. Johann Harimann patiently fielded my endless questions about both teaching and GEO matters, and I miss having him in the corner desk at all hours. And the list goes on: my social and philosophical life has also been enriched by Abdul Ansari, Jason Byas, Mercy Corredor, Kevin Craven, Paul de Font-Reaulx, Guus Duindam, Mariam Kazanjian, Aaron Glasser, Alice Kelley, Calum McNamara, Elise Woodard, Sherice Ng, Caroline Perry, Josh Peterson, Mica Rapstine, Julian Rome, Elizabeth Beckman, Francisco Calderón, Lindy Ortiz, Sarah Sculco, Jonathan Sarnoff, Joe Shin, Ian Fishback, Alvaro Sottil de Aguinaga, Valerie Trudel, Margot Witte, Yixuan Wu, Sophia Wushanley, Glenn Zhou, Sara Aronowitz, Anna Edmonds, and Mara Bollard—and undoubtedly others who I have failed to list.

I am also grateful to my fellow psychology grad students. After I kept taking classes with them, Julia Smith and Wilson Merrell quickly became two of my first friends in psychology. I'm grateful to them for making me feel welcome in the psychology department and for their friendship over the years. Martha Berg basically taught me how to do psychology; she welcomed me into Ethan's lab, showed me the ropes of running studies, and essentially taught me R and stats. I've been lucky to have her as a labmate, friend, and co-author. The members of the Emotion and Self-Control Lab have provided feedback, insight, and encouragement over the years (with special shoutouts to Micaela Rodriguez, Chayce Baldwin, Izzy Gainsberg, Walter Sowden, and Darwin Guevarra), as have those in the Language and Cognition Lab Group (especially Ella Simmons and Valerie Umscheid). The current third year cohort—Desiree Alibar, Imani Burris, Julisa Lopez, and Ariana Munoz-Salgado—welcomed me into all of their first-year things when I was an awkward third year bopping between departments. Yeonjee Bae has been a fantastic collaborator and friend, and Esra Ascigil spent endless time in office hours teaching me stats. Kristi Chin, Soyeon Choi, Clint McKenna, Susannah Chandhok, Nadia Vossoiughi, Irene Melani, and Maggie Meyer have all also been great colleagues over the years. Finally,

thanks to my research assistants: Meriel Doyle, Sophia Katz, and Annalee Miklosek (who is, as I write this, working on coding data from the fourth chapter in this dissertation—Thank you Anna!!).

Some of the most valuable experiences of my graduate career have come from the time I've spent working on various pre-college philosophy initiatives, especially the Ethics Bowl and Philosophy with Kids. Jeanine DeLay is a force of nature. I have loved working with her on the Ethics Bowl and various other projects through A2Ethics—a remarkable organization that our department is lucky to partner with so frequently. I had the pleasure of working with Shelly Venema's Saline ethics bowl team for three years, and I learned a lot from watching her work with her students. Kassia Massey was kind enough to welcome us into her classroom for Philosophy with Kids, and is also a fantastic game night host (and terrifying Jungle Speed player). I am so thrilled to have reconnected in the past two years with Alex Chang, as both a friend and a co-organizer. Together, Alex and I have applied for grants, run Philosophy with Kids programs, taught workshops, travelled across the country, given conference talks, laughed a lot, and consumed too much wine and coffee. I'm grateful for her friendship and her professional collaboration, and I can't wait to see what she does next. And of course, thanks to all the elementary and high school students who have let me talk philosophy with them over the years.

My time in Ann Arbor would not have been the same without the ultimate frisbee community. The bonds that form with teammates have always held a special place in my heart; teammates see you at your best and your worst, and they lead to friendships that may otherwise never have happened. My UM Flywheel teammates are truly the weirdest group of women I've ever met, and I'm lucky to have those weirdos as my friends. Flywheel provided a crucial outlet from the demands of early grad school, and gave me a place where I felt like I belonged. In my years with Rival, I got to push myself to compete at a higher level than I had played at before; I'm so grateful that I got to play with a team that had such a focus on team culture, character, and values—as well as general goofiness and Structured Fun—in addition to high-level play. And to everyone else I've crossed paths within the Ann Arbor/Michigan frisbee world, through tryouts and league and party tournaments: thank you for giving me a real sense of community, friendship, and life beyond academics in this little college town. I won't list everyone's names here, on pain of making these acknowledgements even more rambling than they already are, but I hope you know who you are.

A few frisbee friends deserve special mention. Tia Esposito inviting me to join her house in my second year as a bumbling grad student led to the most fun living situation I've ever had; that year with Liv Perfetti, Phoebe Hopp, Casey Singler, and honorary roommates Janine Kerr and Tina Hanson gave me many of my most joyful memories from Ann Arbor. Tia is a fierce friend and has

done an impressive job keeping our group of Chungi together over the years—and wrangling all of us is not an easy job. I can't wait until we all start our commune. I also want to shout out Tina, in particular, who is one of the bravest and most authentic people I know in pretty much every aspect of her life. Megan Gordon was one of my favorite on- and off-field teammates of all time; she is a joy to play and work and be with. Hannah Henkin and Vivian Chu went out of their way to welcome me to Ann Arbor in my first year, and have been good friends themselves and also responsible for many of the other friendships I've formed in this community. Meg Duffy and Alex Kapiamba (aka Bubbles) have been down-for-anything, keep-the-good-times-rolling friends who turn absolutely anything, from Argus writing days to morning sprint workouts to movie nights, into a fun time. Annie Mei, Hannah Gannon, Nina Janjic, and Sylvia Gisler have remained close friends despite our geographic distance. And last but certainly not least, Tracey Lo went from being a coach I was intensely intimidated by to one of my closest friends. TLo has kept me laughing and singing and thinking through cross country road trips, international travel adventures, long hikes, grueling track workouts, frisbee tournaments, food adventures, game nights, bubble tea walks, and so much more. I'm more grateful for her friendship than I can possibly put into words: she is someone who I respect so much, who I've had the most serious and the most absurd conversations with, who I love to eat and exercise and adventure and just chill with, and who cares very deeply about the people and things she's committed to. My time in Michigan would not have been the same without her.

Many friends beyond Michigan have been an important part of my life throughout these years as well. My Syzygy family has remained a central part of my life, especially Katie Ciaglo, Claire Rostov, Emma Nicosia, Emily Kampa, Ellen Jacobus (throw badly but always throw together!!), Claire Thallon, Maddie Preiss, Eliza Skoler, Sylvie Polonsky, Sheff Sheffield, Elaine Sundburg, and Chessy Cantrell; having this network of friends across the country has meant I'm hard-pressed to travel to a conference where I can't find someone to grab a coffee with. Connor Kasch has been one of my best friends since our first year of college; I'm grateful for continued friendship, long phone call walks, and travel adventures. I've known Joule Voelz since I was three years old and I told her she couldn't join me and my mom while we were gardening. Joule is a rare lifelong friend with whom I feel connected even when we go a long time without talking; it has been such a joy to celebrate each other's successes over the years. All my love to all of these people, and all the others who I've failed to mention here. Writing all of these names has me feeling truly overwhelmed, in the best of ways.

Two other important women in my life are my godmother Maureen Griffin, and my aunt Ann Soter, and long phone calls with both of them have kept me sane throughout grad school. Maureen

has always felt like an extra parent in all the best ways. She's tough and hilarious and empathetic and insightful, and a model for what I hope to be. She's always been one of my absolute favorite people to talk to, either just to chat and catch up or when I'm struggling with a hard decision. Ann is someone I feel I've been able to connect with even more in recent years; I'm grateful for our all conversations and her fun attitude towards life. My love also goes out to Mike Pederson, my godfather and the best storyteller I know, and my godbrother Conall Pederson, who I've had the privilege of watching grow into a real-life, taller-than-me adult over the years.

Finally, thanks to my partner Mason White, who has been endlessly supportive of me throughout my graduate career. Through all the late nights and ridiculous work loads and high stress deadlines and unreasonable travel plans (for school and frisbee), Mason has been there and has never made me feel like I should do anything short of throwing myself fully into my goals and commitments. Thank you for listening to me rant (about the good things and the bad ones), for baking late night cookies, for making me laugh, and for being a source of calm through the whirlwind of chaos that I bring everywhere I go. I love you. And of course, thanks to our sweet pup Nessie, the most opinionated dog in the world, who has perhaps singlehandedly kept me sane during the pandemic.

There's a lot to worry about in the world right now. My entire dissertation has been written during the Covid pandemic. A week ago, the Supreme Court overturned Roe vs. Wade. This summer has already seen one of the hottest Junes on record. It genuinely seems like democracy is crumbling around us. The academic job market is a disheartening prospect. Trying to write papers and run studies through all of this can sometimes feel like a Herculean effort. Nevertheless, writing out the names of all the brilliant, passionate, loving people in my life helps me feel some genuine hope. All my love to you all.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

This dissertation contains a philosophical project and a psychological project. Together, they explore two central themes, and the relation between them: (1) doxastic control and the ethics of belief, and (2) the moral and epistemic import of close personal relationships. The philosophical project (Chapters 1 and 2) concerns a central puzzle in the ethics of belief: how can we make sense of apparent obligations to believe for moral or practical reasons, if we lack the ability to form beliefs in response to such reasons? I draw on empirical work in emotion regulation to make progress on this problem of doxastic control. The psychological project (Chapters 3 and 4) concerns the role of relational closeness in moral judgment: though empirical moral psychology has traditionally focused on judgments about anonymous strangers, I contribute to a growing body of work showing how personal relationships can dramatically affect moral reasoning.

Chapter 1, "Acceptance and the Ethics of Belief," develops an empirically plausible and mechanistically detailed account of *acceptance*, the attitude classically characterized as "taking a proposition as a premise in practical reasoning and action." I argue that acceptance centrally involves preventing a belief from playing its characteristic role in guiding cognition, reasoning, and action, that this centrally involves a "cognitive gating" operation, and that this view gains empirically plausibility by its analogy to well-studied strategies in emotion regulation. Ultimately, I defend acceptance as *doxastic response modulation*. I propose that this account holds promise for addressing puzzles in the ethics of belief—a domain plagued by the central theoretical challenge of our inability to believe for non-evidential reasons.

Chapter 2, "Reframing Epistemic Partiality," applies my account of acceptance to a specific debate in the ethics of belief: the *epistemic partiality* debate, which asks whether we sometimes ought to believe against the evidence regarding our friends. Though compelling, the partialist view has been plagued by serious objections. I argue that the debate has been focused on the wrong doxastic attitude: recasting our duties of friendship as duties of *acceptance*, rather than belief, satisfies the partialist intuitions, without falling prey to the varied objections against the view.

Chapter 3, "What We Would (but Shouldn't) Do for Those We Love" builds on prior work demonstrating that people say they are far more likely to report a distant other, compared to a close

other, who commits a serious moral transgression. Across four studies, I demonstrate that people not only say they *would* protect close others more than distant others, but also that they say it is *morally right* to show such partiality towards close others. Furthermore, I show that people say that they would protect close others more than they think they should—suggesting that moral decisions involving those closest to us may be a context in which people are particularly likely to fail to do what they think is right.

Chapter 4, "How Relationship Affects Adolescents' Decisions to Report Moral Transgressions," investigates how 6th-9th graders respond to the transgressions of close versus distant others. Given the social importance of peer relationships in adolescence, the role of relationship is central to understanding this stage of moral development. I show that adolescents—like adults—are more likely to report distant others who transgress than close others, and more likely to report serious moral transgressions than minor ones.

INTRODUCTION

This dissertation consists of a philosophical project and a psychological project. The philosophical project centrally develops a psychological profile of the attitude of acceptance, connecting classic work in epistemology to cognitive scientific research on emotion regulation, and arguing that this account can help us make progress on puzzles in the ethics of belief. The psychological project explores how relational distance affects moral reasoning, specifically testing how children and adults respond differently to the moral transgressions of close versus distant others.

I introduce each component in more detail below. But first, let me offer a comment about the relationship between the two halves of the dissertation. These lines of research are somewhat distinct: they focus on different questions, in different disciplines, developed with two different sets of primary advisors. Nevertheless, there are uniting themes in both content and high-level methodology. Both projects take seriously the question of how close interpersonal relationships shape the moral landscape. The empirical projects investigate how close versus distant relationships actually influence people's moral judgments, and the philosophical projects ask, in part, how close relationships ought to influence how we believe and regulate our mental lives.

But beyond this, both projects are united by the meta-methodological view that philosophy and psychology can be enriched when each takes the other seriously. My psychological projects are heavily influenced by my philosophical background: philosophy influences what kinds of questions I find interesting to explore, and also how I go about testing those questions—guiding the hypotheses I seek to test, the way I interpret results, and so on. And my philosophical projects are equally strongly influenced by my psychological training. I strive to root my philosophical theorizing in what we know about how the mind actually works. Accordingly, my philosophy engages heavily with empirical work, and—especially when it comes to ethical questions about our mental lives—I strongly believe that we need to let our normative theorizing about the mind be rooted in empirically adequate descriptive accounts. In the end, even though these projects do explore distinct questions, I do not believe I could have created either half of this dissertation without my experiences and training in pursuit of the other discipline.

With that in mind, let us turn to an overview of the projects. The remainder of this dissertation then consists in the four chapters, and a conclusion in which I highlight some of the future work that stems from my present research.

1. The Philosophical Project

When I was around nine years old, I was getting ready to be confirmed in the Anglican church. Confirmation is a Christian sacrament in which a baptized follower affirms their faith, thus becoming a fuller member of the Church. I have a memory of lying in my bed the night before my Confirmation crying, and my mom coming in and asking me what was wrong. I told her, "I don't believe that a snake could talk!"[1] This is my first memory of feeling guilt over what I took to be an evidentially justified belief (though I would not have used that language at the time): I took my evidence to conclusively support the beliefs that snakes cannot, and never could, talk; but I also felt that I somehow (practically or morally) *ought* to believe that the Biblical story was true.

It strongly seems that we can be morally evaluable in virtue of what we believe. It also seems that sometimes, we have good moral or practical reason to believe something that is not consistent with our evidence. These observations have given rise to the philosophical domain of *the ethics of belief*, which asks about our moral status as believers, and the relationship between how we ought to believe *qua* moral agents and *qua* epistemically rational ones. In recent years, the ethics of belief has seen a revitalization, with many philosophers working to make sense of how we should think about the relationship between epistemic rationality and morality. Though some are tempted to try to explain away apparent conflicts between morality and epistemic rationality, I do not find this strategy compelling in the long run: I think the world often does give us evidence for beliefs that are morally or practically undesirable.[2]

But there is a theoretical challenge that lies at the very heart of the ethics of belief: on the standard philosophical view of belief, *we cannot choose to believe on the basis of non-evidential reasons*. Call this the problem of doxastic control. If I sincerely promised you one million dollars to believe that Obama was still president, you clearly have strong reason to do so—and yet, it just doesn't seem like you have the capacity to form a belief on the basis of that reason. That's just not how beliefs work. Even if we

[1] In reference to the snake that tempted Eve into eating from the Tree of Knowledge of Good and Evil, in the Old Testament.

[2] Of course, we *do* frequently make mistakes in our reasoning, and there are lots of cases where we actually would be rational to believe the morally or practically desirable thing! But I do not think that this will capture every case, and I do not think trying to show why every apparent case of tension can be rationally explained away is an especially fruitful philosophical strategy. It's much more interesting and challenging, I think, to tackle the conflict head on.

take a Permissivist view of rationality, on which evidence can warrant a range of rational sets of beliefs or credences, there will surely still be cases in which what it would be morally or practically good for us to believe is simply not supported by our evidence. But how are we supposed to reconcile the problem of doxastic control with the strong sense that we can have moral and practical reasons to believe in some cases?

My friend Eduardo Martinez recently gave me some advice that a former advisor had once given him: if there's something that keeps annoying you about a literature, that's a good place to dig in philosophically. The problem of doxastic control, and its role in the ethics of belief, was that thing that kept cropping up for me. It is this problem that motivates the philosophical projects presented in this dissertation (and, I hope, much of my future research). The literature varies in the extent to which it engages seriously with the problem of doxastic control, but a fair number of authors seem willing to set it aside for the purposes of argument. This strikes me as a mistake. I think that our normative projects in the ethics of belief will be more fruitful to the extent that they are built upon a descriptively adequate philosophy of mind, and are sensitive to the kinds of control capacities we do and do not have over our beliefs.

A further odd (from my perspective) feature of the ethics of belief and doxastic control literatures is that they are both surprisingly unempirical. Though the question at hand involves descriptive claims about our psychological capacities, much theorizing about doxastic control has occurred "from the armchair," without seriously engaging with psychology or cognitive science. On the one hand, this is not entirely unsurprising—pinning down a psychological/cognitive scientific notion of belief is actually rather tricky, and there is significant work to do in understanding how psychological uses of the concept relate to philosophical ones (and how both of those relate to the folk concept). But on the other hand, a central part of my project in this dissertation and in future work is to propose that we can make significant progress in understanding doxastic control by taking seriously empirical research on the kinds of control we do and do not have over other kinds of mental states, such as emotions and thoughts.

The first two chapters of my dissertation are an attempt to begin to make good on some of these worries: to take the problem of doxastic control seriously within the ethics of belief, to use empirical research to make progress on the problem, and to show that this empirically informed approach can help us better understand both the descriptive and normative landscapes at hand.

1.1. Chapter 1: Acceptance and the Ethics of Belief

My first chapter opens with the worry presented above: that we often seem to have moral or practical reason to believe something inconsistent with our evidence, and yet that making sense of such cases runs headlong into the problem of doxastic control. I argue that a promising avenue for making sense of such cases is to appeal to *acceptance*, rather than belief.

Belief, on the standard philosophical view, involves taking proposition *p* to be true in response to the truth of *p*. It can come in degrees of strength, and it operates functionally as our default guiding background, serving as the basis for inference, reasoning, deliberation, planning, intention, action, and so on. Beliefs are thought to be (rationally or psychologically) responsive only to evidence: information that bears on the likely truth or falsity of *p*; we cannot choose to believe something in response to non-evidential reasons. Acceptance, in contrast, has been characterized by prior philosophers as "taking a proposition as a premise in practical deliberation and action" (see e.g., Bratman, 1992; Cohen, 1989, among others). Compared to belief, acceptance is described as more clearly under our direct voluntary control, responsive to practical and moral reasons, and deployable in specific contexts or as a more general policy. Bratman describes acceptance as *departing from* our default cognitive background of belief: we accept when we rely on something *other than* belief to guide our reasoning, deliberation, and action.

I seek to develop a detailed and empirically plausible psychological profile of acceptance—an account of what actually goes on in the mind of an accepting agent. Taking the classic characterization as a starting point, I argue that acceptance centrally involves a cognitive gating operation in which an agent blocks a target belief state (or set of belief states) from having their characteristic downstream effects on reasoning, inference, deliberation, action, and so on, and restructuring those downstream effects in accordance with the accepted proposition. I suggest that this account makes salient several psychological features of acceptance: acceptance can be cognitively effortful, it can involve mere gating or gating plus substitution, and it is best understood as a temporally extended series of mental acts over time. I argue that this account gains theoretical and empirical plausibility through its analogy to well-studied strategies in emotion regulation. Specifically, I propose that acceptance involves applying the cognitive mechanisms at work in emotional response modulation to belief states. To motivate this, I describe various structural similarities between emotions and beliefs. Most crucially, response modulation strategies in both domains allow us to regulate the characteristic downstream role of a targeted mental state (emotion or belief), in response to moral or practical considerations which are not themselves the right kind of reason on which to *form* the underlying mental state.

I propose that having this detailed psychological profile of acceptance on the table leaves us better positioned to appreciate the role acceptance can play in the ethics of belief. We might distinguish between two goals for my account of acceptance:[3]

1) The Empirical Burden: Provide a psychological profile of acceptance that is detailed and *empirically adequate*—that is, that could plausibly be realized in human minds, consistent with what our best evidence from psychological and cognitive science tell us about how the mind works.
2) The Normative Burden: show that acceptance can get us what we want out of puzzles in the ethics of belief.

My primary focus in this chapter is the Empirical Burden. However, I suggest that the view holds significant promise for the Normative Burden as well.

1.2. Chapter 2: Reframing Epistemic Partiality: Acceptance and the Cognitive Work of Friendship

My second chapter applies the account of acceptance developed in Chapter 1 to a specific debate within the ethics of belief: the *epistemic partiality* debate, which considers whether we sometimes ought to believe against the evidence in matters concerning our friends. I take this project to be a useful test case for the ability of my account of acceptance to meet the Normative Burden. To the extent that acceptance succeeds in resolving the puzzle of epistemic partiality (and I argue that it does succeed), I take it to be a promising sign that acceptance as I defend it has a prominent role to play in the ethics of belief more generally.

Partialism is the thesis that we can owe our friends *belief against the evidence*: that at least sometimes, what we ought to believe (about our friends) *qua* friend is different from what we ought to believe about them *qua* epistemically rational agent.[4] Though compelling, partialism has faced a variety of serious objections. Some of those objections worry that important goods of friendship require us to be responsive to the evidence about our friends: some roles of friendship (like giving advice) might depend on accurate assessments of our friends; friendship seems to involve loving our friends for who they really are; and it seems that authentic friendship involves believing well about

[3] Thanks to Tez Clark for suggesting this breakdown.

[4] Of course, we often believe better about our friends than a detached observer might because we have *better evidence* about our friends. However, the cases of interest here are ones in which the evidence really does seem to justify an unfavorable belief about our friends. The partialist thought is thus that *even* in case where the evidence about our friends does not look good, we might owe them more favorable beliefs, on the basis of reasons of friendship.

our friends because we are attuned to their good qualities. Another pressing (and by now familiar) worry is the problem of doxastic control: that we cannot believe on the basis of reasons of friendship.

I propose that this debate has been focused on the wrong doxastic attitude. What's at stake with epistemic partiality, I argue, are not obligations of belief, but rather of *acceptance*. I show that acceptance can handle all of the varied objections raised against partialism. We can accept in response to practical and moral reasons of friendship (thus handling worries about doxastic control), and acceptance makes no demands of systematic irrationality towards our friends (thus handling worries about honesty and authenticity). I then argue that acceptance captures other positive features of the epistemic landscape of friendship, and in doing so I flesh out my account of acceptance more richly. Acceptance accounts for the restructuring of diverse psychological processes in response to reasons of friendship, and captures the diachronic profile of commitment, effort, and mistakes in the cases at hand. Ultimately, I think acceptance gets us what we want out of the partialist thesis, without falling prey to the many kinds of objections that have been raised against it. Accordingly, I propose that acceptance is an example of the kind of cognitive work that friendship can demand.

2. The Psychological Project

The second half of my dissertation continues with the theme of personal relationships and morality. The next two chapters describe a series of psychological studies investigating how social relationships shape moral judgment. In particular, they focus on how our responses to moral wrongdoing change based on our relationship with a transgressor.

For decades, the vast majority of empirical research in moral psychology has primarily studied people's judgments regarding anonymous strangers. Philosophers often prefer thought experiments devoid of contextual and personalizing details as a valuable tool in normative ethical theorizing. However, if our goal as psychologists is to understand how people actually make moral judgments and decisions in the real world, at some point we need to move beyond abstracted thought experiments and get into the weeds of how various contextual and social factors affect moral reasoning. This need for more contextualized moral psychology has become apparent in recent years (e.g., see Schein, 2020).

One powerful and salient social-contextual factor in moral judgment is our relationship to the people involved in a moral scenario. A growing body of work has begun to demonstrate that social relationships can powerfully shape moral cognition (e.g., Everett et al., 2018; McManus et al., 2020; Waytz et al., 2013; Weidman et al., 2020). It should hardly come as a surprise that people reason differently about a moral scenario involving their sister versus one involving their dentist; nevertheless,

until recent years, the particular ways in which social relationship affected moral judgment had been quite understudied.

One domain in which dramatic relationship effects on moral reasoning have been found is in how people respond to the moral wrongdoing of others. Weidman et al. (2020) presented participants with a series of vignettes in which they imagined witnessing some of their closest relationships (e.g., sibling, best friend) and some of their most distant acquaintances (e.g., mail carrier, distant coworker) commit minor and serious crimes—and they asked participants whether they would report the wrongdoing to an authority figure. They found that participants were far more likely to protect (i.e., not report) close others, compared to distant others, who transgressed—and that this effect was even more dramatic for serious crimes than minor ones. This striking finding has been robustly replicated across various conditions and contexts in our lab.

The work I present in the second two chapters builds on these findings. I explore how what people say they *would* do compares to what they think it would be *morally right* to do, and I begin to investigate the developmental trajectory of this preferences towards close others who transgress. Thanks are owed to Martha Berg, Ethan Kross, and Susan Gelman, my co-authors on the following studies.

Chapter 3: What We Would (but Shouldn't) Do for Those We Love: Universalism versus Partiality in Responding to Others' Moral Transgressions

Do people think it's *morally right* to respond differently to the moral transgressions of close (versus distant) others? On the one hand, both laypeople and philosophers often think of ethics as an *impartial* project, in which one ought to treat all people equally. On the other hand, close relationships are central parts of our life projects, and seem to generate special moral obligations. Prior studies exploring the effect of relationship on responses to others' transgressions have only investigated people's *predictive* judgments about how they *actually would act* in these scenarios; no work had tested how people think they *ought* to act, and whether relationships *should* influence their responses.

In the third chapter, I explore this question across four studies, using Weidman et al. (2020)'s paradigm. In each study, participants were presented with eight vignettes in which a close or a distant other was seen committing a serious theft-based moral transgression. Participants were asked to imagine a police officer approaching them, and to decide whether to report what they had seen. Across these studies, I test two competing hypotheses:

1) The Moral Universalism Hypothesis: predicts that people will say they morally should treat people *equally*, regardless of their relationship to a transgressor.

2) The Moral Partiality Hypothesis: predicts that people will say it's morally right to respond *differently* to close versus distant others who transgress.

Importantly, these hypotheses are tested against the background of prior research demonstrating that people say they *would* be more likely to protect a close other transgressor than a distant other. Thus, we were also exploring whether participants were consistent in what they said was right, and how they predicted they would act.

In Studies 1a and 1b, half of the participants were asked whether they *actually would* report the transgressor, and the other half were asked whether they *should* report. In Study 2, all participants were asked what they both would and should do in every vignette. Across all three studies, participants said they both would *and should* protect (i.e., not report) close others more than distant others. However, the difference between protecting close and distant others was greater for "would" judgments than "should" judgments: that is, people indicated that they would protect close others more than they thought they should. In Study 3, we tested whether this discrepancy reveals a genuine inconsistency between what people say they would and should do—or whether it reflects a folk view that what one *morally* should do does not determine what one should do *all things considered*. Participants were assigned to one of four question types: they were asked to indicate either what they actually would do, ideally should do (the language used in previous studies), morally should do, or overall should do. As in the preceding studies, participants said that they both would and (all kinds of) should protect close others more than distant others. Furthermore, participants said that they would protect close others more than they morally/ideally should protect them—and *also* more than they *overall should* protect them.

Together, these studies strongly support the Moral Partiality Hypothesis: participants consistently said that relationship influenced how they *morally should* respond to the transgressions of close others. They also revealed a consistent discrepancy between how participants said they actually would respond to such scenarios, and what they thought was morally right. This suggests that moral decisions involving close others may be a domain in which people are particularly likely to fail to do what they think is right.

2.1. Chapter 4: How Relationship Affects Adolescents' Decisions to Report Moral Transgressions

My fourth and final chapter explores how the effects of relationships on responses to others' moral transgressions change across development. In this study, I investigate how adolescents (6^{th}-9^{th} graders) respond to the high and low severity transgressions of close and distant others. This chapter bridges two literatures which have seen little prior contact: the developmental psychological literature

on children's responses to moral transgressions, and the emerging social psychological literature on how relationship affects adults' moral judgments.

Adolescence is a time in which peer social ties and friendships become increasingly important (Berndt, 1982; Hart & Carlo, 2005). Yet while there is a rich literature on how children and adolescents respond to the transgressions of others, to my knowledge no prior work has systematically manipulated relationship to examine its effects on reporting transgressions.

This chapter reports a study in which we adapt the methods from Weidman et al. (2020) to suit an adolescent sample, modifying the vignettes to take place in a school setting, with a teacher as the authority figure to report to. A large sample (N = 913) of adolescents were asked to think about either their best friend from school, or a distant student in another class, committing high- or low-severity thefts, and asked whether they would report the theft to an inquiring teacher. Results showed that participants were less likely to report close others than distant others, and less likely to report low-severity thefts than high-severity ones. Unlike in prior adult work, there was no relationship by severity interaction: that is, the effect of relationship on reporting decisions was no different for high vs. low severity thefts. In a second study, reported as a supplement, we tested whether the severity by relationship interaction that has been previously documented in adults (Berg et al., 2021; Weidman et al., 2020) replicated with current college students presented with the school-based vignettes. In this undergraduate sample, we again found main effects of severity and relationship—and, consistent with prior adult work, we also found a significant relationship by severity interaction: the gap between reporting close and distant others was greater for high-severity thefts. Though this undergraduate sample provides some evidence for a developmental change regarding this interaction effect, there is a significant demographic confound between our adolescent and undergraduate samples; accordingly, some caution should be taken in comparing the results of these studies.

Chapter 1

Acceptance and the Ethics of Belief[5]

1. Introduction: A Puzzle about the Ethics of Belief

We sometimes seem to have compelling moral or practical reasons to believe something unsupported by our evidence. For example, an athlete might perform better if she believes she will defy the stacked odds and win her race; someone might feel morally compelled to believe her friend's innocence despite evidence of his guilt;[6] or one might regard belief in a social stereotype as evidentially justified but nevertheless morally undesirable.[7] The apparent tension between practical and evidential reasons for belief crops up in a range of philosophical debates.[8] In each we must grapple with a central theoretical challenge: on the standard picture of belief, *we cannot choose what to believe*. In particular, we cannot choose to believe something unsupported by our evidence on the basis of practical or moral reasons. Rather, belief is "involuntary"—a psychologically immediate reaction to our evidence. The specter of doxastic involuntarism hovers in the background of many puzzles in the ethics and pragmatics of belief: how can we make sense of moral or pragmatic duties, responsibilities, or obligations of belief,[9] if we lack the requisite ability to control our beliefs in response to these non-evidential reasons?[10]

[5] I am grateful to Peter Railton and Chandra Sripada for their extensive discussion and detailed comments on multiple versions of this paper. Additional thanks are owed to Maegan Fairchild, Mica Rapstine, Adam Waggoner, Brian Weatherson, Renée Jorgensen, Corey Cusimano, Sarah Buss, Alex Madva, Susan Gelman, and Ethan Kross for their feedback or discussion, and to Aliosha Barranco Lopez, Henry Schiller, Tez Clark, and Caitlin Mace for excellent conference comments. Versions of this paper were presented at the Michigan Graduate Student Working Group, the Michigan Candidacy Seminar, the 2021 Princeton-Michigan Metanormitivity Workshop, the 2021 Austin Graduate Ethics and Normativity Talks, the 2022 Southern Society for Philosophy and Psychology, the 2022 Pacific American Philosophical Association Meeting, and Athena in Action 2022; thanks to those audiences for their questions and discussion.

[6] E.g., see Keller (2004) and Stroud (2006).

[7] E.g., see Basu (2019b) and Begby (2013).

[8] Including, though certainly not limited to: the debates around moral and pragmatic encroachment (see Jorgensen Bolinger (2020) for a thorough overview), doxastic wronging (Basu 2018; Basu and Schroeder 2019), epistemic partiality (e.g., Keller, 2004; Stroud, 2006; see also Arpaly and Brinkerhoff, 2018; Kawall, 2013), and how we should believe in and about others and ourselves (e.g., Morton and Paul, 2019; Paul and Morton, 2018).

[9] Henceforth, I will refer to doxastic *duties* for simplicity.

[10] A number of authors have provided accounts of how we might make sense of epistemic deontology in the face of involuntarism: e.g., see Flowerree (2017). Hieronymi (2008; 2006), Shah (2002), Steup (2012), Weatherson (2008), among others. However, though some of these arguments propose that we do have enough control for deontological concepts to apply, they generally do not claim that we have the control to believe for entirely non-epistemic reasons.

Two overarching strategies are available for trying to make sense of the possibility of a robust ethics and pragmatics of belief. We might find a way to make belief itself do the necessary work—for instance, by allowing the threshold for (justified or rational) belief to vary with the moral stakes,[11] by explaining away any apparent conflict between morality and rationality, or by adopting a view of belief that involves an active psychological commitment.[12] Most have taken this route. But there remains an alternative, less-explored approach: we could instead find a belief-adjacent mental attitude that plays the role of allowing us to respond doxastically to non-evidential reasons. Such an attitude must be sufficiently belief-like to earn its proper place within the ethics of belief while still being legitimately responsive to moral and pragmatic reasons, capable of bearing the weight of the ethical and pragmatic puzzles at hand, and it must be *psychologically realistic*—i.e., actually part of our empirically adequate cognitive ecologies.

In this paper, I pursue the second route. I anchor my account in the existing literature on *acceptance*—which has historically been distinguished from belief as more clearly under our voluntary control.[13] But the picture I develop goes substantively beyond existing accounts of acceptance, by paying special attention to developing a *mechanistically precise and empirically adequate psychological profile of acceptance*. I argue that it is only once we have a psychologically rigorous understanding of acceptance that we can properly appreciate its ability to bear the moral and pragmatic weight we want it to—thus making it deserving of a prominent role in a psychologically realistic ethics of belief.

This paper proceeds in four stages. I start by canvassing the features of belief that make getting a grip on moral or pragmatic doxastic duties so challenging, then turn to existing accounts of acceptance as an alternative (§2). I argue that acceptance is a promising candidate to play an important role in the ethics of belief—*if* we can develop a psychological profile of the attitude that is psychologically realistic. I undertake this project, first proposing an operational account of acceptance as centrally involving a *cognitive gating function* (§3). I then suggest that we can find empirical evidence for the existence of the mechanisms needed to instantiate this gating function by looking to the science of emotion regulation, ultimately leading to my characterization of acceptance as *doxastic response*

[11] This is the strategy favored by certain encroachment theorists.

[12] Something like this "active endorsement" picture of belief seems to be at work in much of Basu's discussions of doxastic wronging, for instance (e.g., Basu 2018), though she does not explicitly defend such a view. See also McKaughan (2007) for an overview of this "active endorsement" account.

[13] Most prominently, this distinction has been developed by Bratman (1992), Cohen (1989; 1992), and Engel (1998) in epistemology. See also Frankish (2007) and Van Fraassen (1985) for related discussion.

modulation (§4). With this detailed account in hand, I return to the question of whether acceptance can bear the weight of puzzles in the ethics and pragmatics of belief (§5-6).

2. Belief and Acceptance

There are several features of belief that make the attitude troublesome for the practical and moral domains,[14] and which have motivated past theorists to look to acceptance.

2.1. A Basic Account of Belief

Most basically, belief is the propositional attitude of taking some proposition *p* to be true, in response to the truth of *p* (or a degree of confidence that *p* is true). Beliefs operate, functionally, as our *default cognitive background* in deliberation and planning: they provide us with a representation of how the world is, thus guiding us through it.[15] They serve as the basis for inference, reasoning deliberation, intention, and action—and do so non-inferentially, without requiring our conscious supervision or intervention.[16]

To usefully serve as our default cognitive basis for moving through the world, beliefs must be responsive to and reflective of how the world actually is. Beliefs are thus characteristically evidence-responsive: they change when we encounter new information that bears on the likely truth or falsity of the proposition in question.[17] We need not oversee the formation of our beliefs in response to evidence; rather, beliefs are in general spontaneously evidence- and experience-sensitive—they form more or less automatically in response to the evidence we encounter.[18] Further, we are in many domains rather good at forming accurate beliefs in response to the world's evidence. Under normal circumstances, and given a reasonably cooperative and evidentially accurate information environment, our belief-forming mechanisms tend to deliver us rather reliable results (especially concerning the

[14] Of course, characterizing belief is not a trivial philosophical task, as there are divergent views about what it means to believe a proposition. I try here to highlight features of belief that I take to be both widely enough accepted by philosophers, and characteristic of the classic evidentialist/involuntarist picture of belief, so as to be a defensible starting point.

[15] I borrow and elaborate on this characterization from Bratman (1992), integrating it in particular with ideas from Railton (2014).

[16] This feature is crucial for understanding the role of belief in our cognitive ecology. The set of things we believe (at least implicitly, though certainly not occurrently) is indefinitely large, and in practical deliberation and reasoning we rely on many beliefs as background premises without consciously entertaining all of them. We could never navigate the world if we had to consider every belief we relied on; we simply lack the time and cognitive resources this would demand. So while we might be able to make many of our beliefs explicit and occurrent, we needn't do so in order to rely on them.

[17] E.g., see Shah (2003); Shah and Velleman (2005).

[18] Of course, we do sometimes step in and deliberate about what to believe in response to a complex body of evidence, but such cases actually represent a very small portion of our total belief-formation experiences.

local, observable world). This is a deeply important feature of our cognitive systems: our beliefs would be a poor guide to navigating the world if they systematically failed to represent the world accurately.[19]

This picture of belief highlights the features that make conceiving of an ethics of belief so notoriously tricky. First, belief is thought to be, on the standard evidentialist picture, rationally and psychologically determined by our evidence: information that bears on the likely truth or falsity of *p*. Pragmatic and moral considerations do not themselves give us justification to think something is (un)likely to be true. Thus, most philosophers maintain that pragmatic and moral reasons cannot rationally be direct reasons for belief as such. Further, because belief is (either conceptually or psychologically) constrained by evidence in this way, and because beliefs generally respond spontaneously to evidence when all goes well, belief is not under our direct voluntary control:[20] we cannot choose to believe for non-evidential reasons.[21] Herein lies the central problem for any robust ethics or pragmatics of belief. Insofar as we accept some kind of "ought implies can" principle, it becomes quite difficult to explain how we could have a doxastic duty to believe without or against our evidence—even in cases where we appear to have compelling practical or urgent moral reason for such a belief—if we lack the doxastic capacity to choose to respond to such reasons.

We could try to escape this tension by adjusting our concept of belief. One could reject this evidentialist, involuntarist picture—though it is foundational enough that many would find this an unattractive route.[22] Alternatively, we could focus on the "indirect" strategies we can sometimes use to influence our beliefs (e.g., seeking out or avoiding particular sources of evidence), theorize about amount of evidence needed for a belief to count as rational or justified given the relevant stakes, or try to explain away any apparent conflicts between epistemic rationality and morality or practicality.

---

[19] It is highly plausible that in many instances, the domains where people systematically and persistently form false beliefs involve epistemic environments that are deficient or distorted in some significant way (perhaps in conjunction with pernicious or disordered motivational factors). When beliefs are false, it often becomes difficult to reason and act on the basis of those beliefs, as the world will continue to push back on the believer. The well-functioning, non-disordered cognitive agent will often find their minds changed by the world, in the end. Of course, this is not true in every case, and there certainly are people who manage to maintain highly unjustified beliefs in the face of significant counterevidence. It may also be easier to cling to such beliefs when they are more abstract and less amenable to everyday evidential support or lack thereof—e.g., the average person's beliefs about the origins of the universe will receive far less pushback than their beliefs about their singing abilities. Yet we should be wary of focusing too much on these anomalous cases at the expense of realizing just how well our belief systems do in general adapt to the evidence they are given.

[20] See Alston (1988) and Williams (1970) for foundational defenses of this claim.

[21] E.g., see Bratman (1992) on Williams: "If I could just choose what to believe at will I could do so independently of whether the content of that belief is true. But if I know of an attitude of mine that is not shaped by a concern with the truth of its content it seems I could not regard it as a belief" (p. 4).

[22] Another possible route is to explain away any apparent conflicts between rationality and morality or prudence. I suspect many (myself included) will find it implausible that there are *never* conflicts between these domains.

But the first of these may be unreliable (if the evidence for the undesired belief is strong, manipulating our beliefs may be both unsuccessful and epistemically suspect) or self-undermining, the second offers little action-guidance to an agent caught in a moral-doxastic dilemma and wades us into the murky waters of encroachment, and the last seems unlikely to succeed to dispel every possible case.

Happily, there is another available strategy: we can look for a belief-adjacent attitude that might allow us to be legitimately doxastically responsive to non-evidential reasons. A promising candidate for such an attitude comes from the distinction that has previously been drawn between belief and *acceptance.*[23]

2.2 A Promising Alternative: Acceptance

Philosophers have sometimes appealed to acceptance in cases where we want to reason and act on the basis of a proposition we do not believe—when we have reason to prevent our beliefs from playing their default guiding role. Bratman (1992) characterizes acceptance as what we take for granted in a practical context; similarly, Cohen describes accepting as "to have or adopt a policy of deeming, positing, or postulating that p… as a premise in some or all contexts" (1989, p. 368). While we may in many cases believe the propositions we use as premises in practical deliberation, *we need not*—we can accept something we do not strictly speaking believe. Acceptance thus allows us to intervene on the operations of that default cognitive background of belief.

Acceptance is a propositional attitude: to accept *p* is to take *p* as a premise in practical reasoning and action. Acceptance differs from belief in several key ways. Perhaps most centrally, acceptance is said to be under our direct voluntary control: it is an attitude we can choose to adopt towards a proposition (Bratman, 1992; Cohen, 1989; 1992). Choosing to accept *p* is said to be a conscious mental act, rather than a psychologically immediate response to a body of evidence. Accordingly, acceptance lacks the distinctive connection to evidence that characterized belief: acceptance does not entail any specific confidence in the likely truth of the proposition in question. We can accept a proposition directly on the basis of practical or moral reasons. For example, a lawyer might accept that her client is innocent on professional grounds and use that as a premise in her reasoning and action regarding her case, even if she believes based on the evidence that her client is guilty. There is nothing inherently

[23] My discussion will focus specifically on the *epistemic* conception of acceptance. As Fleisher (2018, p. 2652 fn4) and McKaughan (2007) note, there are a number of other kinds of acceptance discussed in the philosophical literature, including in philosophy of language, philosophy of science, and literature on metacognition. There may be some systematic differences in how acceptance is conceptualized across domains.

epistemically, psychologically, or rationally suspect about accepting *p* for practical or moral reasons.[24] Further, someone can accept a proposition in some contexts but not others—another way it differs from belief, which is characteristically context-independent.[25]

These existing discussions give us a high-level account of acceptance: a propositional attitude in which we rely on *p* in practical inference and planning, which is under voluntary control, can be reasonably responsive to pragmatic and moral considerations, is not bound by the same evidence-dependent rational norms as belief, can be selectively deployed in specific contexts, and does not depend on whether we believe *p* to be true. The properties I have highlighted enjoy general consensus from the major authors who have written on acceptance (e.g., Bratman, 1992; Cohen, 1989, 1992; Engel 1998; Van Fraassen, 1985 among others), despite some important differences in their accounts.

One difference between existing accounts is worth highlighting, however. Some accounts of acceptance, such as Cohen's (1989; 1992), concerns *all* cases of taking *p* as a premise in practical deliberation, regardless of whether *p* is believed by the agent. While Cohen frequently discusses cases in which acceptance and belief come apart to illustrate their different properties, there is not a fundamental difference on such accounts between cases where acceptance and belief diverge, and where they agree. In all cases, whatever we use in practical reasoning is what we accept; what we believe is an orthogonal question.[26]

Bratman's account, in contrast, identifies a substantive difference between cases where acceptance and belief converge and diverge. Because Bratman characterizes belief as our default cognitive background, acceptance is the mental act that allows us to adjust, or intervene on the operation of, this default background. If our ultimate goal is—as I propose it should be—to develop a psychologically plausible account of acceptance, Bratman's account provides the more useful starting point. In characterizing acceptance as the adjustment or intervention on the normal operations of our default cognitive background of belief, he captures that there is something unique going on when we

[24] There may in fact be cases where our acceptances have negative epistemic consequences—for instance, by leading us to act in ways that contribute to poor evidence-gathering practices. However, this need not always be the case when we accept.

[25] It would be both irrational and highly psychologically odd for one to believe (for example) that Mercury is the closest planet to the sun on Tuesdays, but not to believe this on Thursdays. The same problem does not arise for acceptance: a lawyer might accept her client's innocence in the courtroom, but not at brunch with her friends.

[26] In a similar vein, Stalnaker (1984) characterizes acceptance as treating a proposition as true—but takes this to be a broad category that includes belief as sub-kind, along with presupposition, postulation, assumption, and other nearby attitudes. In virtue of this breadth, Stalnaker's notion of acceptance is sufficiently different from the kind I am concerned with that I will not discuss it further here.

have reason to accept something other than what we believe. From a psychological perspective, this is crucial: we don't need a new account of how reasoning and action are guided when we do not wish to accept against or beyond our evidence, because we already have a perfectly well-functioning cognitive construct to play this role: belief. It is specifically the cases where we choose to adjust or intervene on the operation of that default cognitive background that need elucidation.

Accordingly, I follow Bratman in appealing to acceptance specifically in cases where we reason and act on the basis of something other than what we believe. Accepting is what we do when we intervene on or adjust our default cognitive background: when we prevent the targeted belief from playing its usual role in deliberation and action.

2.3 Is Acceptance up to the Task?

Our hope in turning to acceptance was that it might offer us an alternative, belief-adjacent attitude that can help make sense of the strong intuition that there are cases where we can and ought to be doxasitcally responsive to non-evidential reasons. Can it indeed fill this gap?

Two desiderata for such an attitude are that it is sufficiently belief-like so as to be properly recognized as a part of the ethics of belief, but also that it can be responsive to moral and pragmatic reasons. The high-level account characterized by Bratman and others meets these criteria: we can, on this view, choose accept something we do not believe directly on the basis of moral and practical reasons. This is a promising start.

But there remain two further desired features. First, we want to ensure that the candidate attitude can bear the weight of the ethical and pragmatic puzzles at issue. Many puzzles in the ethics of belief are quite morally significant—for instance, some have argued that in a structurally unjust world, agents can sometimes find themselves with bodies of evidence that rationally justify prejudiced beliefs.[27] If our proposal is going to be that the agent who takes such a belief to be rationally justified but morally undesirable ought to *accept* an alternative proposition, acceptance had better be a sufficiently robust attitude to fill these moral shoes. Second, the attitude must be psychologically realistic, and our account empirically adequate. That is, we ought to have good reason to suspect that the attitude we propose really does exist in agents like us, and that we can and do use it in the contexts at hand.

The existing high-level account of acceptance does not, as it stands, clearly satisfy these two criteria. Armed only with the descriptions of "departing from our default cognitive background" and

[27] See Begby (2013), Basu (2019), among others, for discussion of such cases.

"taking a proposition as a premise in reasoning and action," we lack a rigorous and empirically adequate understanding of the psychological profile of acceptance and the cognitive mechanisms used in its deployment.[28] In the absence of such an account, we also lack a precise understanding of what it actually takes for agents like us to accept, and what differentiates acceptance from other nearby attitudes like supposition—and this, in turn, leaves us ill-equipped to clearly judge whether acceptance is indeed up to bearing the weight of pressing moral and pragmatic doxastic duties.

Here, I take up the project of developing a profile of acceptance that is psychologically detailed and empirically plausible. Once we have this in hand, we can return to the question of whether acceptance can indeed fill the role we need it to in the ethics of belief.

3. An Operational Profile: The Gating View of Acceptance

So, what kinds of psychological operations are involved when an agent accepts?

To accept, in the sense developed here, is to intervene on the default cognitive background of belief, preventing a belief (or set of beliefs) from playing its usual (spontaneous, non-inferential) guiding role in reasoning and action. We will seek to do this when we have some reason not to rely on this default background: when there is some target underlying belief state (or set of beliefs)[29] which doesn't serve our moral or practical aims.

Acceptance must involve identifying all the places where that target belief is activated and operating in deliberation and planning–requiring us to engage in a cognitive monitoring operation. We must then work to block any of the usual inferences, actions, and other downstream effects caused or licensed by that belief state. Finally, we will (at least in some cases; more on this later) need to replace the target belief with the accepted proposition. Doing so will require the agent to extend any inferences and update plans that depended on that belief accordingly.

When we only want to accept in limited contexts, this will not seem so difficult. Accepting that it will rain this afternoon, against my evidentially justified belief that it probably will not, will be fairly straightforward: I might remind myself to pack an umbrella, to make backup indoor lunch plans, or to park my car closer to my office than usual. Yet it will be more difficult to accept in a more context-general way, in part due to the lack of a clearly delineated domain in which the proposition

[28] Throughout this paper, I use the term "cognitive" in the broad sense, understanding it to include, e.g., affective and motivational components.

[29] Throughout this discussion, I will refer to a "target belief." However, it is likely that in many cases, an accepting agent targets not just one single proposition, but a set of related propositions. The view I develop here is entirely compatible with this; references to acceptance and belief of a single proposition should be understood as a shorthand for the set of relevant propositional attitudes.

will be relevant and involved in our deliberations. Imagine a parent who believes their child is lying about something, but decides to accept what they say—not just in some specific context, but in general. If the parent and child live together and are heavily involved in each other's lives, this belief is likely to be activated in a diverse set of inferences, motivations, and decisions. Thus, the parent will have to *monitor* their cognition for all the places where the unwanted belief is likely to be activated in reasoning, decision-making, and other cognitive processes (e.g., attention and motivation). Then, having identified these instances, they must *intervene* to prevent their belief from playing its usual functional role as the default basis for reasoning and action.

We can thus understand the heart of this profile of acceptance as involving a kind of *cognitive gating operation*. This process intervenes to block a particular belief state from having its usual downstream effects and playing its usual role across deliberation, inference, and action. Call this the Gating View of acceptance:

> GATING VIEW OF ACCEPTANCE: Accepting a proposition *p*, when the agent believes that not-*p*, centrally involves preventing the target belief that not-*p* from having its usual downstream effects in deliberation, inference, and action (and restructuring these downstream effects in accordance with *p*).

This gating results in changing the agent's inferential and deliberative landscape, by blocking certain inferences that would be licensed by their beliefs, and blocking certain inhibitions imposed by those beliefs. Thus, acceptance both prohibits some options that would be available to the agent in virtue of what she believes, but also permits moves that would not be available under belief's guidance.

The Gating View is entirely consistent with the existing high-level account of acceptance provided by Bratman and others. But fleshing out acceptance in this way already gives us new insight into key psychological features of acceptance that were not obvious with only the high-level account in hand.

First, acceptance is *cognitively effortful*. One must first be vigilant for all the various places where the target belief(s) may be involved in reasoning, cognition, and action, and intervene to block it from playing its usual role. To do this fully would involve not only monitoring the inferences and decisions that may depend on that belief state, but also anticipating the ways it may affect other cognitive functions–such as motivations, desires, attentional patterns. Further, in cases where one is substituting or adding an additional accepted proposition, one must then also update all these functions accordingly. These operations will involve executive processes which require significant mental resources (of which we have a notoriously limited supply); this is consistent with well-established

evidence about monitoring and intervention processes across cognitive psychology. Thus, we can easily imagine that one might be prone to accidentally relying on their default belief in situations where she is distracted, or engaged in some other cognitively demanding tasks, and thus lacks the available cognitive resources to successfully monitor and intervene upon her default cognitive background.

Second, the Gating View reveals that acceptance is not a one-off action—though classic accounts tend to characterize it this way. Rather, acceptance commits us to a series of specific cognitive actions to be executed for as long as we are committed to accepting *p*. It is thus more precisely understood as a *temporally extended sequence of specific mental acts*, than as a single one-off action. Accepting is something we can choose to do, but so choosing involves committing to a pattern of acts over time. This invites us to ask about the extent to which an agent succeeds in accepting *p*, and evaluate this in terms of how consistently successful she is at identifying contexts where the underlying belief is a premise in her reasoning processes, and how frequently she succeeds at overriding this default belief with the accepted proposition. We also begin to see room for an element of skill or habituation here–ideas that have not frequently been considered in the context of acceptance previously.

Third, the Gating View helps make sense of the relationship between *mere gating*—which only requires the agent to block the undesirable belief's downstream effects—and *gating with substitution*: blocking the target belief's downstream effects, plus the additional component of replacing the unwanted belief with a new accepted proposition. Gating with substitution involves an additional operational step, and so mere refusing to accept and accepting change the inferential landscape in slightly different ways. Mere gating just blocks the inferential permissions and inhibitions enabled by the targeted belief(s). Gating with substitution blocks those, and also issues new permissions and inhibitions resulting from the substituted proposition. Alternatively, acceptance may be used to "shore up" a commitment to a thought-to-be-likely proposition so that we can rely on it in practical deliberation, in cases where an agent is uncertain about whether *p*, or thinks *p* may be likely but doesn't quite *believe* it. (One might want to describe such cases as *addition*, rather than substitution.) This reveals an important feature of the view: that we can apply the gating operation to *whatever* the underlying belief state is—including states of uncertainty—if we have some reason for not wanting that underlying belief state to be the one that guides our reasoning and action.

The Gating View provides an initial operational profile of acceptance—one which accords with existing high-level descriptions, but which also provides us with new insight into features of acceptance that were not obvious with only the high-level account in hand. A natural next step is to ask about the view's empirical plausibility. Beyond being intuitively plausible and cohering with

existing epistemological accounts, do we have good reason to believe that this gating process is something that might actually be realized in human minds? Does the operational process presented here actually accord with any existing phenomena studied by psychologists, that could implement this profile in our cognitive systems?

4. A Mechanistic Profile: Acceptance as Doxastic Response Modulation

The answer, I think, is yes. I propose that empirical research on emotion regulation provides support for the kinds of mechanisms at work in the Gating View, thus lending it strong empirical plausibility. Acceptance, I argue, can be understood as the doxastic analogue to emotional response modulation.

4.1. Response Modulation in Emotion Regulation

The study of emotion regulation–how people can and do control the experience and effects of (often maladaptive) emotions–is a thriving field in psychology. This vast body of research explores a wide variety of emotion regulation strategies, used spontaneously and acquired through intervention training, and compares their effectiveness across contexts and populations. These different strategies are often categorized by the point in the emotion generation and experience process that they target. Most simply, we can distinguish between antecedent-focused and response-focused strategies (Gross, 1998a).[30]

Antecedent-focused responses seek to affect or prevent the formation or elicitation of the target emotion. This includes strategies such as situation avoidance, distraction, and reappraisal techniques. In contrast, *response-focused strategies* seek to regulate the physical, verbal, behavioral, and cognitive consequences characteristic of an emotion, after the emotion has already been elicited, and are thus also referred to as "response modulation" (e.g., Gross and Levenson, 1993; see also Gross, 1998b; Koole, 2009; McRae, 2016 for reviews). One well-studied example of response modulation is expressive suppression, in which one tries to inhibit the facial, verbal, or bodily expression of a felt emotion–such as keeping a neutral face upon witnessing something disgusting or getting upsetting news, or resisting an urge to flee in fear or lash out in anger.[31] Expressive suppression involves

[30] Later accounts more precisely divide strategies into five categories: situation selection, situation modification, attentional deployment, cognitive restructuring, and response modulation (Gross, 1998b; McRae, 2016). Further, because emotions are temporally extended mental processes, the line between categories is in practice somewhat blurry. However, since I will be discussing only response modulation in detail, the courser two-category distinction is sufficient for present purposes.

[31] Although regulation is often discussed in the context of negative emotions, people can regulate positive emotions as well. For instance, someone trying to keep a neutral face and hide excitement upon learning that they were accepted into a prestigious school, or stifling laughter in response to a funny video, are examples of expressive suppression for positive emotions (Gross and Levenson, 1993).

monitoring one's verbal and behavioral responses for the characteristic expressions of felt emotions, and preventing oneself from enacting those expressions.

These strategies can be used in specific one-off instances, such as when trying to suppress a yelp of fear when watching a scary movie, or repeatedly, as a long-term strategy for frequently arising emotions (Gross and John, 2003). In both cases, these response modulation strategies are demanding on executive processes (Franchow and Suchy, 2015, 2017; Gyurak et al., 2012; Niermeyer, Franchow, and Suchy, 2016; Richards, 2004; Richards and Gross, 1999): is cognitively effortful for people to deploy them over time, and they interfere with performance on other executively demanding tasks. People vary not only in how much they tend to rely on suppression techniques in everyday life, but also in how successful they are in using them to chronically regulate emotion. Crucially (and in partial explanation of the characteristic effort), these response modulation strategies do nothing to directly affect the initial generation of the emotion; their target is instead the downstream action tendencies (and in some cases, cognitive effects) that characteristically flow from the emotion's activation.

4.2. Doxastic Response Modulation

I propose that acceptance involves applying the cognitive mechanisms at work in emotional response modulation to belief states.[32] The response modulation process as developed for emotion regulation accords with key components of the Gating View.

Most centrally, acceptance–like emotional suppression–does not directly seek to alter the underlying psychological base state (belief, in this case) itself. Rather, both seek to prevent the characteristic downstream consequences of the activation of the state in question. The Gating View characterizes acceptance as seeking to prevent the targeted belief from playing its usual role in deliberation, reasoning, and action. To do this is, essentially, to *suppress* the normal downstream effects of an activated belief–just as emotional suppression seeks to suppress the downstream effects of a felt emotion.

There are of course differences between the characteristic downstream profiles of belief and emotion. However, these differences are not as deep as they may first appear; both kinds of states can give rise to the same broad categories of downstream effects. Just as strong emotions often give rise to facial expressions, verbal reactions, and other physical action tendencies, so too do beliefs. Encountering evidence that is starkly incongruent with belief-generated expectations may lead to

[32] Those who endorse accounts on which belief is a form of confidence or trust, such as Railton (2014) and perhaps Schwitzgebel (2002), and insofar as confidence and trust are affective states, might simply say that acceptance is a specific kind of emotional response suppression.

experiences and physical and verbal expressions of surprise. Successful acceptance–for instance of a friend's innocence, against a belief in their guilt–may involve preventing oneself from letting out a surprised "oh!", raising one's eyebrows, or doing a noticeable double-take when encountering new evidence that *does* seem to speak strongly in favor of their innocence. Just as suppressing a fear response may involve resisting the urge to turn and run, so too does accepting a friend's innocence involve resisting belief-motivated action tendencies such as not trusting them (behaviorally or psychologically) with sensitive information.

Beyond these expression and action tendencies, the characteristic downstream profile of belief also includes a significant cognitive component as one of its primary functional roles: beliefs normally serve as our premises for reasoning and inference (our "default cognitive background"). Theorizing about acceptance thus requires us to focus on the suppression of these cognitive and inferential tendencies. On its face, this may appear to be a point of difference with emotional suppression, where empirical work has focused far less on suppressing cognitive effects. Yet emotions certainly do affect our reasoning processes, for better or worse. Although we think of downstream cognitive effects as being more central to the role of belief than to emotion, it would be a mistake (and in conflict with decades of empirical research on how emotions effect reasoning) to suggest that emotions do not in fact have significant cognitive consequences. Thus, fully suppressing an anger response surely involves not only preventing particular facial expressions, but *also* recognizing when our anger influences our reasoning processes–perhaps driving us to draw more negative conclusions about someone than we would if we were not angry. While this kind of cognitive response may be quite difficult to prevent–and it may well be the case that brute suppression is a particularly ineffective strategy for controlling the downstream cognitive consequences of emotion–this difficulty does not itself signify a deep difference between the broad kinds of downstream effects that belief and emotion can have.

We thus have an initial case for a parity between the primary function of emotional response modulation and acceptance: blocking the downstream effects of an activated state. The phenomenological profiles are similarly analogous. In suppressing both emotion and belief, there will be an experience of blocking, gating, working hard to prevent the usual responses from emerging–whether those are facial expressions, action choices, verbal reactions, or reasoning patterns. In the cases of suppressing cognitive and inferential responses, there will be an experience of trying to monitor and identify what parts of the reasoning processes have been influenced by the belief or emotion, and to adjust those accordingly. Further, the phenomenology of *effort* is prominent across

both domains, and this accords well with the research cited above indicating that emotional response modulation is cognitively demanding.

I'll highlight three further dimensions of similarity between emotional and doxastic response modulation. First, we noted that the Gating View allows us to piece together the relationship between mere gating and gating with substitution: the Gating View accommodates the observation that in some cases we merely seek to gate a targeted belief state, while in others we seek to actively substitute a replacement proposition as well. A similar set of options exists the emotional domain. Sometimes, an agent merely wants to mask an emotional reaction and suppress the characteristic expressive responses, such as maintaining a neutral face when feeling disgust. Other times, she wants to actively replace the usual disgust response with an alternative response–such as a display of (apparent) enjoyment. Someone feeling disgust towards a friend's poor cooking may not only wish to show a neutral face, but to actively seem as though she is enjoying the food.

Second, it is characteristic of both emotional response modulation and acceptance that our deployment of these strategies is not entirely limited by the features that rationally or psychologically constrain the formation of the underlying psychological states. In other words: the very point of these regulation strategies is to provide agents with a way to override the default cognitive and behavioral roles of mental states that arise in response to particular kinds of stimuli. Our affective systems respond to certain kinds of stimuli–danger, injustice, impurity–with specific kinds of emotions–fear, anger, disgust. When these emotional processes are well-attuned, the activated emotions will be *fitting* to their objects; there will be certain perceived features that appropriately give rise to the emotions in question. Emotion regulation strategies allow agents to exert some control over the downstream effects of those emotions, when we have practical or moral reason to do so, even when the activation of the emotion was reasonable and fitting. Similarly with acceptance: doxastic response modulation gives agents a way to exert control over the effects of beliefs in our reasoning and deliberation, even when those beliefs are evidentially warranted and well-formed, when we have practical or moral reason to do so. In both domains, these response modulation processes allow agents to be responsive to practical reasons which would not themselves be the right kind of reason (descriptively or normatively) on which to directly form the underlying base state. Response modulation (emotional and doxastic) thus expands our agential capacities without sacrificing the proper functioning of the underlying system.

Finally, in developing the Gating View we observed that accepting may be less difficult in a limited domain or context, but can become quite challenging over a wide range of contexts or long

period of time. This too tracks well with the cognitive profile of emotional response modulation. While it requires some cognitive effort to mask one's anger or sadness in a one-off context, it becomes far more difficult to do this repeatedly over time and across a wider range of contexts. In the domain of emotion, relying on expressive suppression in daily life compromises general cognitive performance (Franchow and Suchy, 2015), is associated with negative affect (Brans et al., 2013), and has even been thought to contribute to depression and anxiety disorders (Campbell-Sills et al. 2006; Kashdan and Steger, 2006; Sperberg and Stabb, 1998;) and stress-related symptoms (Moore, Zoellner, and Mollenholt, 2008). In the context of belief, it is not difficult to imagine that it is far easier to accept a proposition that is only relevant in a specific time or place (such as that it will rain this afternoon) than a chronically relevant one (such as that your friend is innocent, or that your cooking skills are superior)—and that the greater set of contexts in which one seeks to accept, the more difficult it may be. Further, emotion regulation theorists generally agree that suppression is a rather maladaptive regulation strategy (Gross, 1998b; Lynch et al., 2001), associated with a host of negative psychological and social effects (Butler et al., 2003; Richards and Gross, 1999; Richards, 2004). Regulation strategies that seek to avoid the elicitation of the emotion in the first place (e.g., reappraisal strategies) are more sustainable and beneficial in the long run (Gross, 1998b; Richards and Gross, 2000).

This too may hold a lesson for the doxastic domain: if you anticipate committing to accepting something in the long term and across contexts, it may be psychologically easier and more cognitively efficient to try to get yourself to believe the proposition in question instead–by searching for new evidence or reinterpreting old, generating further explanations, etc.–or to simply avoid conditions that make the belief occurrent and activated. However, actual belief manipulation strategies face several challenges: they take time, they cannot be guaranteed to work (sometimes the evidence just isn't out there, and of course one of the starting premises of this entire project is that we cannot always get ourselves to believe something merely out of prudence), and any such belief-manipulation strategies, even if ultimately successful, may be epistemically suspect. Thus, acceptance is a crucial belief-management strategy available to us in the meantime, or in situations when we do not actually want to interfere with our well-functioning belief-forming mechanisms.[33]

[33] It is worth noting that an agent who successfully accepts something over a long period of time may in fact end up altering their beliefs, without that being their intended goal. Insofar as acceptance will influence the agent's patterns of behavior, planning, reasoning, and inference, the accepting agent is likely to alter her evidence-gathering practices in a way that the non-accepting agent is not—thus delivering her a different body of evidence than the non-accepter. This means that long-term acceptance may still result in different patterns of beliefs. But importantly, direct belief manipulation is not the target goal of this process.

These similarities across the emotional and doxastic domains give us compelling reason to think that we can posit response modulation as a plausible mechanism for acceptance. The profile of response modulation aligns neatly with the key features of both the Gating View and the high-level existing accounts of acceptance, and also has the benefit of being a well-established mechanism in the domain of emotion, thus lending it significant empirical plausibility–a crucial element for the present project of moving beyond high-level epistemological theorizing about acceptance and into the realm of empirically adequate philosophy of mind.

5. Integrating Acceptance into the Ethics of Belief

We now have on the table a detailed psychological profile of acceptance, including a clearer picture of what kinds of mental operations it employs and the nature of the cognitive mechanisms involves. This leaves us better situated to explore a remaining question about acceptance: can it bear the weight of puzzles in the ethics and pragmatics of belief?

With only the high-level account of "taking a proposition as a premise in practical reasoning," we might have worried that acceptance felt thin or unsatisfying compared to the robustness of belief—a kind of cognitive pretense, or "acting as if." When the motivations in question feel practically or morally urgent—e.g., when they involve the success of one's future endeavors or the effort to be a good friend—we might have felt hesitant at "settling" for "mere acceptance."

But by building out the profile of acceptance as we have here, I believe we can see that acceptance *is* robust enough to stand up to the moral and practical import of many puzzles in the ethics of belief. Acceptance on my view involves a serious kind of commitment to regulating one's cognition and action over time—there is nothing thin or "hand-wavy" about setting out to do this in the long run. The response modulation mechanisms involve significant cognitive resources, and require cognitive vigilance to identify all the ways in which a target belief affects one's cognition and action. Accepting will often be difficult, and require real buy-in from the agent—and they may well still fail to fully regulate every consequence of their belief. Still, an agent who sincerely tries to accept for some moral reason, and who does a good (even if imperfect) job at doing so, has done something morally significant and praiseworthy. Thus, we do not let an agent off the hook by suggesting that their doxastic duties are ones of acceptance rather than belief.

5.1. Is Acceptance All We Care About?

Even if convinced that acceptance can have a role to play in the ethics of belief, one might still wonder just how much acceptance actually solves. An objector could press:[34] imagine an agent who (we stipulate) entirely succeeds in accepting her friend's innocence, but does nothing to alter her underlying belief that her friend is guilty. Does the accused friend still have any claim to be upset? If she acquires a device that gives her access to all her friend's mental states and thus discovers that she still in fact *believes* she is guilty, would or should she still feel wronged by her friend?[35]

I'll provide two responses to this objection. The first is that my account reveals just how difficult it is to stipulate that an agent entirely succeeds in suppressing her belief. In many real-world cases, there will likely be moments in which the underlying belief state breaks through the agent's suppression efforts–and thus, someone who wants to resist the idea that mere beliefs can (be) morally wrong could locate their complaints about the agent in the places where the undesired belief still influences her cognition and action. This also accords with the fact that the agent herself would likely recognize these breakthroughs as failures, and might feel guilty about them.

The second route is more conciliatory. For all I've said here, it remains a genuinely open question–with room for reasonable disagreement–whether the belief state *itself*, independent of any downstream effects on cognition and action, and independent of any upstream failures of irresponsible reasoning or evidence gathering, can (be) morally wrong. Exploring this question is an important project in the ethics of belief, and settling it falls beyond the scope of this paper. At the very least, however, I believe that having a robust account of acceptance shifts the burden of proof: the proponent of belief *itself* must now give a positive argument defending its distinctive importance in some domain, independently of its downstream consequences in cognition and action.

5.2. Acceptance and Supposition

Another way in which lacking a psychological profile of acceptance may have obscured its theoretical usefulness, is that without a clear understanding of the psychological mechanisms involved, it can be tempting to conflate acceptance with nearby attitudes like supposition. With our new account of acceptance in hand, we are now in a much better position to clearly distinguish acceptance from supposition, thus making acceptance's distinctive role in our cognitive ecologies even more clear.[36]

---

[34] Thanks to Chelsea Rosenthal for doing precisely this, and to Maegan Fairchild for further discussion of this point.
[35] See Basu and Schroeder (2019) for more discussion of doxastic wronging.
[36] Thanks to Chandra Sripada for encouraging me to flesh out this distinction.

The high-level goals of acceptance and supposition are plausibly quite different. Acceptance allows us to intervene on our default cognitive background when we want to reason and act on the basis of something we do not believe; acceptance is taking *p* as a premise in practical deliberation. Supposition, in contrast, is characteristically a kind of hypothesis exploration, mental simulation, or (potentially counterfactual) investigation of "imagine if *p* were true." We have characterized acceptance as involving a kind of cognitive commitment to reasoning and acting on the basis of *p*. Supposition does not seem to involve this same commitment, and is unlikely to be a policy we commit to over time and across contexts. Instead, its aim seems much narrower: we generally suppose for the specific purpose of figuring out what might follow if *p*.

Because of these differing goals, supposition characteristically involves constraining or restructuring our reasoning and inferences, but not our actions, reactions, or any broader set of cognitive mechanisms such as attention and motivation. We have discussed at length how acceptance, in contrast, involves blocking all the various cognitive, expressive, and behavioral downstream effects of the target belief. To see this, consider the different between agent A who accepts it will rain this afternoon, and agent B who supposes it will rain this afternoon. While we expect A to adjust her planning, inferences, and behavior to accord with a rainy afternoon, B does not seem subject to the same demands: instead, we imagine her merely thinking through what might happen if it rains, how things may be different than if it remains sunny, etc. B's aims are specifically exploratory and epistemic, and do not practically commit her to "it will rain this afternoon" in the same way that A's acceptance does. Similarly, an agent who merely supposes her friend is innocent may not be required to suppress a surprise reaction upon encountering new evidence of their innocence, or actively attend to the various weaknesses in the evidence against her. The fact that supposition does not require the agent to prevent all the various downstream responses of a target belief over an extended period of time, and instead merely consider what might rationally follow from the truth of the proposition in question, suggests that the Gating View of acceptance is far more demanding than what is needed for supposition. A supposer needn't monitor her cognition and action nearly as thoroughly as an accepter, and needn't commit to a long-term policy of doing so; the supposer merely needs to restructure her inferences to explore what follows from *p*.

These differences in high-level characteristics may also reflect differences in the lower-level psychological profiles of acceptance and supposition. Though a full treatment of the psychological profile of supposition is beyond the scope of this paper, we can offer some initial thoughts. As a first pass, one could argue for a gating and response modulation account of supposition with a more limited

target: the supposer only needs to gate and suppress the reasoning and inferential processes involving the target belief. This may capture part of the terrain, but it may not be the only available route. Instead, it seems plausible that supposition involves somewhat of a different cognitive profile, centrally involving processes of counterfactual reasoning, hypothetical simulation, and cognitive decoupling–and that these processes are more emphasized than the monitoring, gating, and suppression mechanisms that characterize acceptance. Such processes more closely align with the exploratory goals of supposition.

These considerations suggest another dimension of difference, one which also helps clarify an important feature of acceptance. Following other authors, we have repeatedly claimed that acceptance is under our voluntary control: we can choose to engage the regulatory mechanisms discussed here. But, we have emphasized that these processes are effortful and cognitively demanding, and so agents may often fail in particular instances to regulate their target beliefs. What has not yet been articulated is that beyond this general observation about our limited regulatory capacities, some acceptances will be more difficult than others. Though acceptance lets us act and reason on the basis of propositions that are not supported by our evidence, our ability to successfully accept some *p* is not entirely unconstrained by our evidence. The more dramatic the departure from our overall evidence, the more of our default cognitive background of belief we will have to monitor and intervene on, and the more difficult successfully accepting across contexts will be. We thus start to see the limits of our acceptance-regulatory processes, and appreciate that although acceptance is not rationally determined by the features that determine belief, it is *constrained by*, influenced by, and sensitive to them in important ways.

Supposition, in contrast, seems not to face these same restrictions. We can suppose things that drastically conflict with our evidence and are wildly implausible–even though it would be extremely difficult to actually succeed at accepting these things. I can suppose that there is a hoard of angry elephants outside my apartment, or that the speed of light is a different constant; I can reason through what might follow if either of these things were true. Yet actually accepting these, in the sense that I have argued for in this paper, would be implausibly difficult.[37] The fact that supposition has aims

[37] I noted earlier that *entirely* succeeding at accepting even more reasonable propositions is quite difficult, in the sense of identifying and blocking every single downstream consequence of the target belief. However, we can surely be reasonably successful at this in many cases. But imagine just how difficult it would be to try to gain even a modest amount of success at accepting the presence of a hoard of elephants outside my (American college town) window. Surely actually aligning my behavior and cognition with this proposition would involve paying attention to nothing but my (apparently empty) lawn, calling my friends and the local news, taking pictures with my phone, and so on.

which are less practical and thus does not require me to act on the basis of these claims, and the fact that we tend to suppose in more limited contexts and timeframes, allows an agent to suppose a whole range of things that they may struggle to accept.

This suggests that the more an agent seems to be *regulating*, committing to *p*, and using it to guide *action*, reaction, and a range of cognitive capacities such as attention in addition to inference, the more this will look like acceptance. In contrast, to the degree that an agent seems to have merely exploratory aims and is restructuring only their reasoning and inference patterns with the goal of figuring out what would follow if *p*, it may be more apt to describe them as merely supposing. It may be difficult to know precisely where to draw the line between these attitudes in some cases, especially when describing the psychologies of other people–and some may resist making the distinction at all. But, for those who want to understand how acceptance differs from supposition, I believe considering the characteristic aims and potential psychological profiles of both attitudes offers us a way to do so.

I'll close by briefly considering an alternative way one might have been tempted to distinguish between acceptance and supposition. At the outset, one might have tried to characterize acceptance as having specifically practical (broadly construed) aims, and supposition as having specifically epistemic aims. I think this characterization does not work for the simple reason that I believe acceptance, as I have described it in this paper, *can* be undertaken for specifically epistemic aims. To acknowledge this is to point out that sometimes, being the most successful epistemic agent in the broad sense will involve responding not merely to the considerations of the evidence directly in front of us.

A classic instantiation of this point is a scientist who favors a hypothesis that is simpler or more explanatorily lovely, but which is equally or even less supported by the evidence in hand.[38] In such cases, the scientist may at times accept the proposition, even if they do not strictly speaking believe it–but the motivation for doing so appears characteristically epistemic, insofar as it is connected to a motivation to learn more about the world, build more successful scientific models, etc.[39] More broadly, however, we can imagine an agent who knows she faces an evidential situation

[38] Such cases motivate authors like Van Fraassen (1985) to appeal to acceptance: it allows them to explain how either practical or otherwise non-evidential factors could play a legitimate role in scientific epistemology.

[39] The philosophy of science is one of the domains in which acceptance has received more historic and contemporary attention. Yet while I think acceptance as developed here does have a role to play in scientific cognition, I doubt it tells the whole story. In many cases, I suspect that the kinds of cognitive regulation mechanisms I've argued for may be stronger than what the scientist needs; I am not convinced that reasoning and experimenting on the basis of a working hypothesis always requires the kind of systematic doxastic suppression I have discussed. A more appropriate attitude, in many cases, may be something like Fleisher (2018)'s *rational endorsement*, which focuses on broader norms of inquiry rather than a specific cognitive profile. I similarly suspect that in the philosophy of religion—another domain where

that will be deeply unreliable or otherwise hostile. Though she expects the incoming misleading evidence, she also knows that her belief-forming mechanisms, left to their own devices, are likely to still respond to that evidence, update on it to some degree, or even be overwhelmed by it. Thus, she might seek to regulate her beliefs via the kinds of acceptance mechanisms discussed here, for the clearly epistemic purpose of retaining overall better beliefs, despite her impending evidential situation.

These cases provide just a preview of the many ways in which acceptance can also be used for characteristically epistemic purposes, and thus suggest that contrasting the evidential and the practical is not the most fruitful way of distinguishing supposition and acceptance. There is far more to be said about the use of belief regulation mechanisms for epistemic aims than I have the space to discuss here, and so further theorizing on that topic must be left for future work.

6. Concluding Thoughts

We began this paper by observing that there a recurring tension in puzzles in the ethics and pragmatics of belief: on the one hand, there are many cases in which we appear to have compelling practical or moral reason to believe, but on the other hand, it is classically held that we cannot choose to believe on the basis of such non-evidential reasons. I have proposed that perhaps many of these apparently moral/pragmatic doxastic duties can be reframed as duties of *acceptance* rather than belief. This strategy becomes more compelling when we try to move beyond existing high-level accounts of acceptance and flesh out an empirically plausible psychological profile for acceptance—leading us to the Gating View of acceptance, fleshed out cognitively as doxastic response modulation. I have sought to develop a profile of acceptance that accords well with the high-level account initially presented by authors like Bratman, but also which gives us new insight into characteristics of acceptance that were not apparent with only the high-level account in hand.

This newly enriched account of acceptance helps us to see why the attitude is well-suited to address certain kinds of puzzling problems. Acceptance allows us to be genuinely responsive to practical and moral considerations, and to regulate our beliefs in accordance with these reasons–without forcing us to posit an implausibly voluntaristic account on which we can simply choose what to believe directly in response to these reasons. But acceptance also does not force us to require the doxastic gymnastics of trying to trick ourselves into holding a belief, nor does it force us to have the aim of infringing on our well-functioning belief-forming mechanisms in a way that is likely to be epistemically suspect.

---

acceptance has received more attention—the profile developed here may have some role to play, but that it will fail to capture significant aspects of religious cognition.

Acceptance deserves a prominent role in the development of a robust and psychologically realistic ethics of belief. For instance, perhaps those seeking to defend epistemic partiality in friendship (e.g., Keller, 2004; Stroud 2006) could make progress by appealing to acceptance. Similarly, an agent who holds a racially prejudiced belief may have compelling reason to suppress that belief across a wide range of contexts. Such suppression could be motivated by merely professional or reputational concerns, or by genuinely moral ones, as in the case of someone who has an attitude of *regret* towards her prejudiced belief.[40]

More speculatively, acceptance may help clarify the landscape in other less overtly moral domains of belief. For instance, acceptance may play some role in the kinds of cases that tempt some to appeal to "the power of positive thinking." A very ill person may accept that she will recover, despite strong evidence to the contrary. A runner who "decides to believe" that she will win the race as she steps onto the track may be incorporating acceptance into her pre-competition ritual. Acceptance may have a role in understanding *grit*: in undertaking difficult long-term projects, we may at times need to suppress our beliefs about our likelihood of failure.[41] Conversely, acceptance may play a role in certain kinds of self-deception or self-denial. In each of these cases, acceptance may not tell the whole story, and there may be other kinds of doxastic regulation or propositional attitudes in play– exploring this remains an opportunity for future work.

I'll conclude by highlighting three key takeaways. First, I hope to have vindicated the role of acceptance as an important component of our cognitive ecologies, and a helpful tool in epistemology, philosophy of mind, and the ethics of belief. Second, I hope this present work invites us to expand how we think about doxastic control and the scope of our doxastic agency. I have tried here to shed light on a way in which we can exert significant control over our doxastic states, through the lens of *regulating* of our underlying beliefs, rather than through the traditional lens of choosing what to believe. Finally, I hope to have shown by example that the topic of doxastic control, which has long been primarily the purview of armchair philosophy, can be significantly elucidated with help from the empirical psychological sciences. Ultimately, this discussion of acceptance should be understood not as an ending point, but rather as a starting point: inviting us to both consider what other doxastic

[40] Rapstine (2021) develops this idea of *epistemic agent regret* more fully, building on Bernard Williams's conception of agent regret in the moral sphere. I find the heart of Rapstine's proposal compelling: the idea that we can hold a belief, take that belief to be evidentially justified, but nevertheless regret being a "vehicle" for that belief on moral grounds. Acceptance gives us a resource to do something about our beliefs in such cases, rather than merely resigning ourselves to this regret.

[41] See Morton and Paul (2019) for a thorough philosophical treatment of this topic.

capacities can be understood via empirical investigation, and also to consider what roles belief regulation capacities such as acceptance may play in our epistemological, practical, and moral lives.

Chapter 2

Reframing Epistemic Partiality:
Acceptance and the Cognitive Work of Friendship[42]

1. Introduction

What do we owe our friends, epistemically? This question has been discussed at length in the recent philosophical literature on *epistemic partiality in friendship*—which asks whether friendship can make demands on our epistemic and doxastic lives, and whether those demands of friendship can stand in tension with the norms of epistemic rationality. Various answers have been proposed in response to this question, including that we owe it to our friends to believe in them when they undertake difficult endeavors (e.g., Marušić & White, 2018; Paul & Morton, 2018), to take special epistemic care when we reason about them, to double check evidence that paints them in a bad light (Arpaly & Brinkerhoff, 2018; Goldberg, 2019), to believe well of them, to look for the good in them and not dwell on the bad (Basu, 2019b; Brinkerhoff, forthcoming; Gardiner, mansucript; Keller, 2004, 2018), to trust them (Hawley, 2014), to interpret evidence about them especially charitably and expend extra effort in thinking through matters concerning them (Stroud, 2006; see also Gardiner, manuscript)—and so on.

These proposed obligations fall (roughly) into two categories. The first category contains obligations that target our "upstream" epistemic practices—those concerning, for instance, when and how we inquire into matters concerning our friends, how thoroughly we gather and verify evidence regarding them, how much effort we put into thinking through and understanding situations involving our friends. These practices are surely important features of the epistemic landscape of friendship, and this is where some have located most or all of our obligations in this sphere.

[42] I am grateful to Chandra Sripada, Peter Railton, Renée Jorgensen, and Maegan Fairchild for their detailed feedback on various drafts of this paper, and to Susan Gelman, Ethan Kross, Gabrielle Kerbel, and Zach Barnett for further discussion. Versions of this paper have been presented at the University of Michigan, Duke University, the 2022 Munich Graduate Conference in Ethics, and the 2022 Rocky Mountain Ethics Congress; thanks to those audiences for their discussion, and to Simon Stromer and Helen Han Wei Luo for their conference comments.

The second category contains obligations that look like they require us to *believe against the evidence* when it comes to our friends. For instance, consider the following case:

> CHEATING: Mateo comes to have good evidence that his best friend Shelby may have cheated on her final statistics exam. A reliable teaching assistant insists she saw Shelby staring suspiciously at her inner arm, where she now has a large smudge of ink, too obscured to tell what was written. Shelby has gotten in trouble for cheating once in the past. And Shelby's performance on the exam was quite a bit better than her performance on past exams in the class. Although Shelby told the teaching assistant that she studied extra hard, Mateo has been busy with his own finals and does not have evidence to verify that. **Mateo feels torn: he thinks his evidence suggests that Shelby likely cheated, but he feels he owes it to her as her friend to believe she did not**.[43]

Some authors (most notably, Keller, 2004; 2018; Stroud 2006) have argued that in cases like this one, friendship can make demands on our beliefs: what Mateo ought to believe *qua* good friend is different from what he ought to believe *qua* rational epistemic agent.

My focus in this paper will be this second category of doxastic obligation in friendship.[44] Many have found cases like CHEATING compelling[45] —there are, it seems, situations in which it a good friend will believe against the evidence (and these intuitions cannot be satisfied by appealing only to modifications in upstream practices). But defenses of such obligations face pressing objections, rooted in both the nature of friendship and the nature of belief. Most centrally, they face the worry that we cannot choose to adopt beliefs for moral or practical reasons—a worry which threatens to entirely head off the debate about owing our friends belief against the evidence. This leaves us with a puzzle

---

[43] I focus here on cases where a friend's testimony is not part of the evidence. I agree with Goldberg (2019) that our friends' testimony can provide us with strong (though not insurmountable) evidence for the truth of what they say, especially in high-stakes cases. As Goldberg argues, given reasonable background assumptions from both sides about the importance of maintaining trust in a committed friendship, it is often evidentially reasonable to believe that our friends are telling us the truth in high-stakes scenarios, because they know that to lie would compromise trust and damage the friendship. However, there are many cases where we either do not have access to testimonial evidence from our friends, or where the belief at stake is not one which testimonial evidence can clearly settle. Though Goldberg takes his explanation to apply equally well to cases where a friend's testimony is involved and those where it is not; however, he does not fully explain how his account is supposed to cover the latter—and I think there is more work to do to account for these cases.

[44] Others have labeled this "direct" partialism (Mason, 2021; see also Arpaly and Brinkerhoff, 2018): the thesis that friendship can make demands directly on the content of our beliefs. This is contrasted to "indirect" partialism, which focuses on various practices which may impact the beliefs we ultimately come to have.

[45] Importantly, the specifics of the case are not central here—I offer a case where a friend has been accused of wrongdoing as just as one kind of example that has been frequently offered in the literature. Others might prefer cases involving belief in a friend succeeding at something very difficult and unlikely, or involving beliefs about whether a friend and their partner are a good match—or any other number of scenarios.

which has not been satisfyingly answered by the existing literature: how can we preserve the intuition that we have doxastic obligations towards our friends in such cases, while taking seriously the challenges that have been raised against this picture?

In this paper, I propose a novel solution to this puzzle: what we really owe our friends in these cases is not *belief*, but rather *acceptance*. Specifically, we owe them acceptance in the sense of regulating the characteristic downstream guiding role of a belief in cognition, reasoning, and action—an account developed by Soter (under review).

This paper will proceed as follows. I start by describing and motivating the partialist thesis (§2), and then describe two families of objections that have been raised against it (§3)—objections that we *shouldn't* believe partially about our friends, and objections that we *can't*. I then introduce a specific account of acceptance as doxastic regulation, which has close ties to the familiar architecture of emotion regulation (§4). I show that appealing to acceptance, instead of belief, to make sense of epistemic partialism both accounts for the objections raised against partialism (§5) and also attractively captures other features of the cases in question (§6).

Before diving in, a quick note about "belief." There is reason to suspect important differences between the folk and philosophical conceptions of belief. The philosophical conception of interest is a "thin" evidentialist notion of belief as a mental state that is specifically and immediately responsive to our evidence—a state over which we lack direct voluntary control (especially for extra-evidential reasons), and which represents our confidence in the likely truth or falsity of some proposition *p*. But this may be just a narrow slice of the (admittedly under-investigated) folk notion of belief, which, in contrast, appears to be fairly capacious—for instance, there is empirical evidence that laypeople attribute substantive control over belief (Cusimano & Goodwin, 2019; Turri et al., 2018).[46] It is specifically this thin evidentialist philosophical notion of belief that, as we will see shortly, runs into all sorts of problems in the domain of epistemic partiality. Nevertheless, the solution I ultimately offer of appealing to acceptance rather than belief will, I suspect, allow us to capture these folk intuitions; plausibly, we should interpret folk obligations of partiality as having been about acceptance all along.

[46] Several studies have found that people's attributions of control over belief consistently fell above the midpoint on a seven-point Likert scale (Cusimano & Goodwin, 2019; Turri et al., 2018), and that people attributed more control over beliefs than some other mental states like emotions (Cusimano & Goodwin, 2019). Exactly what people think this control amounts to remains an open empirical question, and an important one to explore. One possible hypothesis is that many of the doxastic-regulatory mechanisms that I defend as part of acceptance are actually captured within the folk notion of belief.

2. The Partialist Thesis

The partialist claim of interest is whether friendships (and other close personal relationships) (at least sometimes) make demands on our beliefs, such that we can owe it to our friends to believe against the evidence in matters concerning them.[47] This suggestion is proposed against the background of *evidentialism* as a dominant view in mainstream epistemology (Feldman & Conee, 1985; Shah, 2006), on which only evidential factors (i.e., information that bears on the truth or falsity of a proposition) are reasons to believe that proposition.[48] The mere fact that someone is my friend is not, in many cases, something that bears on the likely truth or falsity of some proposition about that person. Thus, *partialists*—who endorse the claim that friendship can give us obligations to believe favorably about our friends even in cases where the evidence does not support such a favorable belief—hold that "there are cases in which an agent cannot meet both the highest standards of friendship and the highest standards of epistemic responsibility" (Keller, 2004, p. 330).[49] Most famously, this view has been defended by Simon Keller (2004; 2018) and Sarah Stroud (2006).

In cases like CHEATING, the partialist might claim that Mateo owes it to Shelby to believe her innocence, despite the unfavorable evidential situation. Stroud and Keller propose other examples that share this dynamic as well. In Keller (2004)'s well-known example, Eric's evidence suggests that his friend Rebecca's poetry will likely be mediocre and not impress a literary agent in the audience; nevertheless, Keller suggests that Eric owes it to Rebecca to try to form more favorable beliefs about her work. Stroud (2006) considers a case where your friend Sam recently slept with someone and then never returned her calls, despite Sam's awareness of the feelings she had for him. Stroud assumes that in this case, your evidence of Sam's bad behavior is solid—and yet, she suggests that in such a case you owe Sam "something other than an impartial and disinterested review of the evidence" (2006, p. 504), as his friend. Crucially, the claim at stake here is that we owe it to our friends to believe some

[47] I follow others in this debate in focusing on cases where the belief at stake is a belief about one's friend. However, we might also owe doxastic duties of friendship concerning beliefs about other matters—say a belief about someone or something else important to them.

[48] There is some disagreement about the best way to characterize evidentialism. In the partiality literature, it is sometimes characterized as the claim that what one ought to believe is solely determined by one's evidence. In the literature on encroachment, this stance is called *purism* (Bolinger, 2020).

[49] Though partialism has most often been discussed with reference to evidentialism (and I will follow that trend here), the same tension can arise even on theories of epistemic rationality that allow other considerations to play an important role in standards of belief formation and revision beyond the narrowly evidential: for instance, those that given an important role to explanatory considerations. Against such theories, the partialist thesis can be reframed to claim that friendship will still give us reasons to adopt beliefs that are not necessarily consistent with the beliefs that epistemic rationality would require us to adopt

way about them *despite* the unfavorable evidence[50] (what Mason (2021) labels "direct partialism"); this is distinct from, though consistent with, the claim that we might also owe our friends different "upstream" epistemic practices (such as double-checking our evidence).[51]

These altered doxastic practices are proposed to be motivated by the needs and interests of our friends, not by what is likely to be true. But because the motivation for these practices has nothing to do with making our beliefs likely to be more accurate, this means that friendship is giving us reason to be epistemically irrational by evidentialist lights. Importantly, Stroud and Keller talk not only about partiality being characteristic and constitutive of friendship, but also frequently talk in terms of what friendship *demands* of us, or of what we *owe* our friends.[52] The partialist proposal is thus not merely an evaluative claim about how we should assess people as friends in virtue of their beliefs; rather, this thesis is framed as an action-guiding one, telling us what ought to do to be good friends.

There are several reasons philosophers have found the partialist thesis attractive. First, there is the simple observation that many people feel an intuitive tension between what the evidence suggests we ought to believe and what a friend ought to believe in various cases. There appears to be a reasonably reliable intuition that one's beliefs and mental states can make a difference to the friendship in cases like CHEATING, and others raised in the literature. Further, there is some empirical evidence that this kind of intuition is shared by laypeople: for instance, people indicate that a close friend should believe more optimistically about someone accused of a drug charge than an acquaintance should, and, crucially, that the close friend should believe more optimistically about their friend than they think is permitted by the evidence (Cusimano & Lombrozo, 2021). Insofar as this kind of intuition is reasonably widely shared among both philosophers and laypeople, it seems worthy of serious investigation. Friendship is a domain in which we all have intimate personal experience; thus, we ought to take seriously widespread judgments about what friendship involves.

Moving beyond mere intuitions about particular cases, Keller (2018, pp. 25-26) motivates partialism by arguing that some of the distinctive goods of friendship—including support, having

[50] Partialists acknowledge that sometimes we believe more favorably about our friends than a detached observer because we have more evidence about them. However, the cases of interest are those where we owe our friends epistemic and doxastic duties that are *not* justified by this further evidence—cases where by stipulation, the evidence against our friends does not look good.

[51] Mason (2021) labels this "upstream" partialism "indirect partialism; and Apraly and Brinkerhoff call it "partialism-light." Both take the light/indirect version of partialism to be strictly compatible with the view that there are no practical reasons for belief.

[52] For instance, Keller writes that "the tendency to treat us sympathetically [through their beliefs] is not just one that we think likely to be manifested in our friends, it is one that we can *want* them to manifest" (2004, p. 338; emphasis in original).

someone on your side, and openness to seeing the world through our friends' eyes—require us to sometimes form beliefs based on reasons of friendship, rather than on the evidence. He also notes that what others think of us—especially those with whom we share close relationships—can make a difference to our wellbeing on any account that goes beyond subjective hedonism (Keller, 2018, pp. 20-24). He proposes that we want our friends to believe well about us—even (or perhaps especially?) in cases where the epistemic chips are down and things don't look good for us. In this sense, he argues, we have a stake in what our friends believe about us—even if we don't actually know what those beliefs are.[53] These lines of reasoning point to an even more basic reason in favor of the partialist thesis, which is that it gets something very fundamentally right about friendship: that friendship makes demands not only on how we treat our friends, but also on how we *think* about our friends.

3. The Partialist Puzzle: Some Key Objections

Nevertheless, when we interpret these claims on the standard philosophical account of belief, this partialist thesis runs into several serious problems. The objections that have been raised against partialism fall into two main categories: those that target the "friendship" part of the partialist thesis, and those that target the "belief" part.

3.1. Objections from Friendship

The first family of objections against partialism claim that we *should not* believe partially and irrationally about our friends, because there are important goods and obligations of friendship that require our beliefs to be based on evidence about who are friends really are. There are several versions of this worry.

Let's call the first version the *objection from honesty*. It holds that partialism overlooks the importance of having honest assessments of our friends (see Arpaly and Brinkerhoff, 2018; Kawall, 2013; and Mason, 2021). The first way to motivate this is by pointing out that there are important cases in which having an overly inflated view of our friends would interfere with our ability to fulfill certain roles of friendship. For instance, we frequently seek out our friends' advice before undertaking big life choices or risky endeavors, such as the decision to open a new business. Our friends' ability to give sound advice in such situations depends in crucial ways on their having a sufficiently honest and accurate view of us to form a reasonable belief about our likely chances of success—without this, they

[53] Similarly (though with a broader focus) Basu (2021) suggests that "we care how we feature in the thoughts of other people and we want to be regarded in their thoughts in the right way" (p. 109).

might encourage us to put our life savings into a business that's doomed to fail.[54] Thus, the argument goes, there are salient cases in which holding more favorable beliefs about our friends than would be justified by the evidence alone would actually make us *worse* friends.

Of course, all this shows is that we *sometimes* need honesty from our friends' beliefs; this does not disprove the thesis that other times, friendship will indeed involve evidentially unjustified beliefs. While honest belief might be important when a friend is seeking frank advice, optimistic belief might be important when what they need is support or motivation. The problem here is that once we've adopted an irrational belief in one context, this will make trouble for the cases where honesty and evidential rationality is needed. Beliefs (on the standard philosophical view) are characteristically acontextual and stable across contexts (barring changes in one's evidence). Thus, a friend who succeeds in adopting an irrationally favorable belief about my chances at business success in some context will not (on the standard view) be able to simply discard that belief in a context where evidentially-justified beliefs become crucially important.

There is a deeper dimension to the honesty objection: we want our friends to love us based on *who we actually are*, not based on some fictionalized, idealized version of us. Kawall points out that if a friendship is based on systematic illusion and unjustified belief about someone, we should worry that the friendship is flawed, and "not a love of the friend herself, with her actual character and qualities" (Kawall, 2013, p. 361; see also Gardiner, manuscript). Similarly, Mason (2021) proposes a Murdochian account of friendship on which knowledge of a friend's true character is a central part of friendship. Both authors (but especially Mason) suggest that systematically irrational beliefs about important features of one's friend undermines the legitimacy of the friendship, as they seem to prevent one from loving and relating *to the person themself* as a friend. It is surely right that we want our friends to know and love us for *us*, and not for an idealized fiction of ourselves.[55]

A related objection about the importance of evidentially-grounded belief in friendship is *the objection from authenticity*. This objection worries that believing on the basis of friendship itself is the wrong kind of reason (in a normative or evaluative sense) for thinking well of our friends. Crawford (2019) proposes that to believe favorably about our friends, on the basis of *thinking that is what one ought*

[54] Arpaly and Brinkerhoff (2018, p. 43) and Kawall (2013, p. 360) both worry about such cases. For instance, Kawall discusses cases where—as a result of encouragement from a friend with "positive illusions" about their friend's abilities—someone takes on tasks or goals for which they lack the requisite skills.

[55] Further, being clear-eyed about someone's flaws or bad actions need not always undermine our friendship with them: surely we have all had the experience of fully recognizing a friend's flaws, but loving them anyway. See Yao (2020) for a compelling discussion about the importance and challenge of loving someone with full recognition of their flaws.

to do to be a good friend, is to be *inauthentic* towards our friends. Insofar as we hold positive beliefs about our friends, we should hold them because think those beliefs are rationally warranted: because we are attuned to the good-making features of our friends that justify those beliefs. On this view, we are not being good friends when we believe well about our friends because we think that favorable beliefs about one's friends are the kinds of attitudes we think friends should have; rather, we are good friends when we believe well about them because we notice, attend to, and appreciate the qualities in them that make them deserving of our positive assessment and estimation.

3.2. Objections from Belief

The objections from honesty and authenticity raise crucial concerns about the nature of friendship. But even if we alleviate these worries,[56] there remain pressing objections to partialism grounded in its claim about how we ought to believe.

Perhaps the most obvious objection to partialism is a persistent worry for the ethics of belief: *the problem of doxastic control*. On the standard philosophical accounts of belief, we do not have direct voluntary control over our beliefs; in particular, we cannot choose to believe something unsupported by our evidence directly on the basis of practical or moral reasons.[57] Beliefs are, on this view, not only rationally but also psychologically or constitutively determined by our evidence. But if we cannot choose to believe on the basis of non-evidential reasons (such as moral and pragmatic considerations), and assuming we accept some kind of "ought implies can" constraint on our duties and obligations, then how can we make sense of any claims that there are moral or pragmatic duties to believe something? At its core, this objection retorts that we do not owe it to our friends to believe against the evidence because we *cannot* believe against the evidence for reasons of friendship (see Arpaly and Brinkerhoff, 2018; Goldberg, 2019 (p. 2230, fn. 17)).

One way to try to avoid this problem is to appeal to various strategies we might adopt to try to cultivate favorable beliefs about our friends in situations that demand such beliefs. But this too faces worries: as Arpaly and Brinkerhoff (2018) point out, philosophers are often a bit too happy to throw around the idea that we can cultivate specific mental states in ourselves—but this is actually quite difficult to do with much precision or reliability. And the worry is particularly bad in the case of

[56] We might distinguish between the proposal that we ought to have a policy of believing more favorably about our friends *in general*, versus the proposal that there are some specific cases in which we ought to believe more favorably about our friends, but that in general rational belief is warranted. The objections from honesty and authenticity seem to have far less bite against the claim that we *just sometimes* owe our friends belief against the evidence than against the claim that we ought *in general* believe better about our friends than the evidence warrants—and the latter thesis does indeed seem implausibly strong.

[57] See Alston (1988) and Williams (1970) for particularly foundational defenses of this claim.

belief—for if we alter our epistemic practices with the stated goal of trying to bring about a particular belief in ourselves that we don't take to be supported by our evidence, this processes risks becoming self-undermining (Williams, 1973). So the problem of doxastic control persists.

A final salient challenge to partialism simply denies that the epistemic and doxastic obligations owed to our friends are epistemically irrational. For instance, Kawall (2013) proposes that friendship demands at most a *modest epistemic bias* towards our friends: the beliefs we form in cases like those discussed above may indeed be different than those of a disinterested observer, but this does not yet show that such beliefs are irrational.[58] We might flesh this idea out by appealing to epistemic permissivism,[59] which holds that there is not one unique (set of) belief(s) or confidence state(s) that is rationally required by any given body of evidence (e.g., Kelly, 2013; Schoenfield, 2014). On such a view, there might be room for practical/moral considerations like those of friendship to influence which (set of) belief(s) to adopt from those that fall within the rationally permitted range—thus allowing friendship to act as a reason to adopt a belief without condemning us to irrationality or requiring us to believe against our evidence.[60]

However, I do not think this objection succeeds at discharging the puzzle of partialism, because it always seems open to the partialist to continue presenting cases that further stack the evidential deck against our friend. In CHEATING, for instance, we can introduce further unflattering pieces of evidence against Shelby—and the intuition might still remain that Mateo owes her some kind of doxastic duty of friendship. Our goal should not be to kick off an endless string of case-swapping—first, because this is an unproductive philosophical approach, and second, because our goal here is *not* to develop an account of why Mateo might be evidentially rational in believing that Shelby is innocent after all. Rather, the present aim is to try to make sense of precisely the cases in which friendship does

[58] Relatedly, Paul and Morton (2018) argue that purely epistemic considerations are often not enough to pick out a uniquely best "evidential policy"—a standard that structures our reasoning, thinking, or belief-updating practices—and so there is room for practical and contextual factors to influence the beliefs we adopt. They suggest that the limits of epistemic rationality to determine a unique appropriate evidential policy are especially salient when forming beliefs about agents' future endeavors ("believing in" people).

[59] See also Hawthorne (2014) and Goldberg (2019) for use and discussion (respectively) of this strategy

[60] Another way to develop a version of this objection would be to adopt an encroachment view, on which moral and practical considerations can legitimately affect what it is rational to believe (see Bolinger, 2020 for an overview). However, encroachment is a controversial view (and, as Bolinger argues, there is not one single kind of encroachment thesis). We will be strategically better off if we can develop an account of partialism that does not rest on antecedently accepting encroachment, and I will not wade into the weeds of encroachment in this paper. For more on the relationship between encroachment and partiality, see Goldberg (2019) and Gardiner (manuscript).

seem to come into conflict with epistemic rationality—those cases when we really seem to owe our friends some kind of doxastic duty, despite the unflattering evidence—whatever those cases may be.[61]

3.3. Taking Stock of the Puzzle

We find ourselves stuck with a puzzle. On the one hand, we have the persistent intuition proposed by partialists that we can sometimes owe it to our friends to believe against the evidence in matters regarding them (in addition to whatever other epistemic obligations of friendship we might have); this is a key component of the partialist view. On the other hand, we have a collection of serious concerns about the viability of this view. Such obligations risk being both impossible due to worries about doxastic control, and (even if we could fulfill them) inadvisable given the importance of honesty and authenticity in friendship—concerns which both highlight the importance of believing based on the evidence within friendship. The puzzle is this: how do we account for our intuitions that friends can owe each other specific doxastic duties in cases like CHEATING, without sacrificing important goods of friendship or positing unfulfillable obligations?

As it stands, we lack a clear account that can capture the doxastic duties at stake without falling prey to one or more of the objections discussed above. Prior work in the literature has either retreated to our indirect or upstream duties in response to worries about doxastic control (e.g., Arpaly and Brinkerhoff, 2018; Goldberg, 2019) or else has insisted that we must reject our partialist intuitions to preserve honesty and authenticity in friendship (e.g., Crawford, 2019; Mason, 2021). No account put forward in the literature adequately addresses the motivations and concerns on both sides of the debate. Happily, I think there is a way to capture the motivating spirit of the partialist thesis, while also accounting for the objections that have been raised against it.

4. Acceptance and the Regulation of Beliefs

My proposal is that what's really at stake are not obligations of belief, but rather of *acceptance*—specifically, acceptance in the sense developed by Soter (under review). On this account, acceptance involves a commitment to preventing a belief from playing its usual guiding role in cognition and action. In what follows, I lay out this view of acceptance and show how it can help us account for both sides of the puzzle of partialism: it can give us a story of what we owe our friends when we seem to owe them belief against the evidence, without falling prey to the objections discussed above. It turns out, on this view, that what is at stake when it comes to the doxastic landscape of friendship is

[61] In general, I think we ought to be skeptical of attempts to show that (even permissivist accounts of) evidentialism can *always* explain away conflicts between morality and epistemic (see Traldi, 2022 for one way of defending this idea). The world is an uncooperative place; why should we assume that we can never have morally undesirable beliefs that are nonetheless reasonable and rational given our evidence?

not just the beliefs that we hold regarding our friends—but rather, the roles we do (and don't) allow those beliefs to play in our cognitive economies. What's important is not only *what* we think about our friends, but also *how* we think about them.

4.1. An Account of Acceptance

In domains far removed from epistemic partiality, other philosophers have proposed that in circumstances where we do not want to let some belief play its usual guiding role in structuring our deliberation and action, we can instead *accept* some alternative proposition—where acceptance is characterized as taking a proposition as a premise in practical deliberation and action.[62] In particular, Bratman (1992) describes acceptance as departing from our "default cognitive background" of belief. For instance, a lawyer might believe based on the evidence that her client is guilty, but nonetheless accept the client's innocence on professional grounds and in professional contexts: that is, she might commit to a policy of reasoning and acting on the basis of her client's innocence.[63] Thus, on a Bratman-style account, we can accept when we want to reason and act on the basis of something *other* than what we believe—as belief would normally guide these processes spontaneously and non-inferentially (Railton, 2014).

While belief is (in some sense) involuntary and determined by the evidence, acceptance is proposed to be more clearly under our direct voluntary control. Acceptance is something we can choose to do, and we can undertake this action for practical goals, in specific contexts, and in response to non-evidential considerations. When we choose to accept something we do not strictly speaking believe, we intervene on our default cognitive background to prevent our beliefs from playing their usual guiding role, and we instead commit to acting and reasoning on the basis of whatever we accept.

Elsewhere, I adopt this high-level characterization as a starting point, and from it develop a more precise account of what actually goes on, cognitively, in the mind of an agent who departs from her default cognitive background of belief (Soter, under review). The goal of this account is to characterize acceptance in terms of empirically psychological mechanisms. Here, I outline some of the key features of that account that are important to understand for present purposes.

---

[62] Acceptance has been discussed by a number of authors, including Bratman (1992); Cohen (1989; 1992), Engel (1998), and Van Fraassen (1985). There are important differences between these accounts—for instance, Cohen describes acceptance as *whatever* we take as a premise in practical deliberation, while Bratman describes acceptance as specifically what we do when we want to rely on something *other* than what we believe. Soter (under review) prioritizes the Bratman-style account of acceptance; see that paper for more discussion on the motivation.

[63] Cohen (1992) discusses this kind of example at length.

An accepting agent seeks to reason, deliberate, plan, and act on the basis of something other than what she believes. To do this, she will need to *monitor* her cognition to identify the various ways in which the unwanted target belief is active and influencing her cognition and action. She will then need to *intervene* to block all the usual inferences, actions, and other downstream effects caused or licensed by the target belief—and (where appropriate) substitute in the alternative accepted proposition. This blocking can be described as a *cognitive gating operation*, consisting in preventing the target belief from having its usual downstream role in cognition, deliberation, and action—and then *restructuring* one's deliberative and inferential landscape accordingly. Patterns of thinking and acting that would be licensed (or inhibited) by the underlying belief must now be blocked (or are now permitted) by acceptance. Importantly, these processes can target belief states of any level of confidence, including states of uncertainty.

Two key features of acceptance arise from this characterization. First, acceptance is cognitively effortful and demanding on executive processes, of which we have a limited supply, as it often involves active inhibition of default patterns of thinking and acting. Second, acceptance is not a one-off action, but rather a temporally extended sequence of specific mental acts. Accepting, in other words, involves committing to the gating operation and restructuring over time: for however long and in whatever contexts an agent is trying to accept, she is committed to working to block and restructure the characteristic downstream effects of her target belief in cognition and action. This is something the agent can choose to do: it is under her volitional control. But the control profile here is one of effortful regulation over time.

This account gains both theoretical and empirical plausibility by its operational analogy to well-known strategies in the domain of emotion regulation. One prominent class of emotion regulation strategies are *response modulation strategies*, which seek to regulate the verbal, behavioral, and cognitive consequences of an emotion that has already been elicited (e.g., see Gross, 1998a, 1998b; McRae, 2016).[64] Such strategies target the characteristic downstream effects of an activated emotion state—for instance, facial expressions of disgust or vocalizations of fear—but do not directly target the alteration of the underlying emotion state itself.[65] I argue that we ought to understand acceptance as

[64] These response-focused strategies are classically contrasted with *antecedent-focused* strategies, which seek to prevent the elicitation or activation of an emotion state in the first place, or otherwise intervene upon the generation of the emotion state. This distinction is broad and oversimplifies things, but it is useful for present purposes. See Gross and Feldman Barrett (2011) for a helpful discussion of how the success of the distinction depends on one's background view on the nature of emotions.

[65] These strategies are often labeled "expressive suppression" in the psychology literature.

applying the same cognitive mechanisms at work in emotional response modulation to belief states: acceptance is *doxastic response modulation*.

There are numerous similarities between the gating account of acceptance and the well-studied profile of emotional response modulation, which I spell out in detail in Soter (under review). For now, it is enough to appreciate this at a high level: in both emotional and doxastic response modulation, we rely on cognitively effortful control mechanisms to prevent the characteristic downstream consequences on cognition, deliberation, reasoning, and action—and these suppression efforts do nothing to directly alter the underlying base (emotion or belief) state. In both domains, we must continue to deploy these regulatory mechanisms for as long as a) we are committed to suppressing the unwanted mental state, and b) the unwanted mental state remains (i.e., as long as I am still angry, or still believe that *p*). And crucially, in both the emotional and doxastic domains, response modulation allows us to regulate the mental state in response to practical and moral considerations which are not themselves the right kind of reason on which to form the underlying mental state. In other words, just as emotion regulation allows us to (for instance) suppress our anger in situations where anger is practically disadvantageous *even if the anger was fittingly elicited*, so too can doxastic regulation (i.e., acceptance) allow us to suppress the effects of a belief that is practically or morally undesirable, *even if the belief is well-formed and evidentially justified*. In both cases, we can regulate the underlying states without compromising the proper functioning of the state-elicitation processes.

4.2. Partiality and Acceptance

With this account of acceptance as regulating the characteristic downstream effects of a belief on guiding cognition and action, we are now in a place to begin to unsnarl our puzzle of partiality. My proposal is this: in cases where an agent seems doxastically pulled in one direction by the evidence, and another direction by reasons of friendship, what they owe to their friend is not a duty of belief, but rather of acceptance. That is, it's not that we directly owe it to our friends to *adopt* a belief about them that is inconsistent with our evidence. Rather, we owe it to our friends to *regulate* the usual functioning of the undesirable (from the perspective of friendship) belief in our cognitive economies. We owe it to them to commit to blocking the effects of our (well-formed and evidentially justified) belief and to coming ourselves to restructuring our reasoning and action on the basis of the accepted proposition. In the remainder of the paper, I will make the positive case for this view.

5. Resolving the Challenges from Belief and Friendship

First, I will show that appealing to acceptance to explain our doxastic duties to friends can resolve the challenges for partialism discussed in §3. I then provide several further reasons to think

that acceptance accurately captures what partialists want out of the rich doxastic landscape of friendship.

One set of objections against partialism targeted the "belief" component of the thesis, worrying that we cannot have doxastic duties of friendship because we cannot form beliefs for such reasons, or that we don't need them, because any cases of apparent tension can be resolved with a permissivist framework.

Acceptance straightforwardly meets the challenge of doxastic control. Acceptance provides us with a substantive account of how friends ought to respond when evidential reasons for belief and reasons of friendship seem to conflict—without forcing us to either deny the orthodox view that we cannot choose to form beliefs in response to such practical and moral reasons, or to retreat to indirect methods of belief manipulation which may be at best unreliable, and at worst self-undermining. Acceptance takes seriously the limits of our control over belief formation—but also elucidates the substantive *regulatory* control we have over the role beliefs play in our cognitive economies.[66]

Second, because acceptance allows us to make sense of how we can respond doxastically to cases where the evidence, by stipulation, does not rationally permit a positive belief in our friends, we alleviate the temptation to try to explain away all cases of apparent evidential irrationality in terms of (for instance) permissive standards of belief. It may well be that a "modest epistemic bias," permissivist standards of evidential responsiveness, and altered upstream inquiry practices of (e.g.) evidence-gathering, interpretation of information, and patterns of attention can account for a great many cases where we believe differently about our friends than a disconnected rational observer would. But we need not (and, I think, should not) assume that there can never be dilemmas that pull us between epistemic norms and moral ones. Instead, acceptance—which is answerable to a wider range of reasons than the merely evidential—gives us the resources to explain what we owe our friends in those cases where what seems doxastically demanded by friendship is genuinely outside of what it is rationally permissible to believe based on our evidence.

In short, acceptance avoids the problems partialism encounters when we frame it as a duty to *believe* against the evidence, because acceptance is not about belief formation, but rather about belief regulation. Even if we cannot choose *form* beliefs in response to the needs of our friends, we can choose to regulate their characteristic guiding role in our cognitive economies, in response to the needs of our friends.

---

[66] Though as I discuss in §6.2, the fact that acceptance is under our voluntary control does not mean that we have *perfect* control over it.

The other set of objections against partialism worried about important goods of friendship—e.g., honesty and authenticity—whose realization depends on us believing based on our evidence when it comes to our friends.

Acceptance can handle both versions of the objection from honesty. It accounts for the observation that there are important contexts where having an irrationally favorable belief would make us worse friends, such as when giving a friend advice depends on our having a clear-eyed view of the situation. Unlike belief, which is characteristically stable across contexts, we can choose to accept a proposition in some contexts but not others. Thus, explaining partiality in terms of acceptance gives us an important kind of discretion as friends: we can accept a proposition in cases where doing so would be beneficial for the friendship or our friend (e.g., when they need support and encouragement), but not accept it in cases where being guided by rational belief is more valuable (e.g., when they need clear-eyed advice, or perhaps an honest intervention). Keller himself notes the importance of this context-sensitivity, writing: "good friends treat each other differently under different circumstances, and a good friend often has the skill of being able to discern and respond to her friend's needs, as they are and as they change… and the same goes for belief formation" (2018, p. 28). This feature, so puzzling from a belief-oriented account of partiality, is no problem for acceptance.

In a similar vein, acceptance also allows us to account for worries about how we ought to weigh partialist reasons of friendship against other kinds of moral reasons. Gardiner (manuscript), for instance, raises broader moral worries about partialists' reliance on cases in which a friend is accused of serious wrongdoing. Consider a friend who is credibly accused of sexual assault: even if we might in some sense be a good *friend* if we believe their innocence, we risk serious moral harm to the victim in doing so. The acceptance view gives us the right resources to make sense of these worries. Reasons of friendship are just one instance of the moral and practical considerations that govern acceptance. This means that in deciding whether to accept, we ought to take into account not only whether someone is our friend and whether it would be good for *them*, but also various other kinds of moral considerations—including considerations of justice to others involved. In some cases (such as CHEATING, which is relatively mild), reasons of friendship may dominate. In others—such as sexual assault cases—other moral concerns may outweigh considerations of friendship.

This means that figuring out when and whether acceptance is appropriate will be part of the complex part of both friendship and morality. My goal here is not to defend whether acceptance is, on the whole, the right choice in any (kind of) particular case. Rather, I aim to show that *if* there are

times when considerations of friendship seem to come into conflict with epistemic rationality, than acceptance can handle these cases.

The second objection from honesty worried that partialism is incompatible with seeing, knowing, and loving our friends for who they truly are. Suggesting that in some cases we owe it to our friends to accept against the evidence leaves ample room for such honest acquaintance with our friends. The acceptance account recommends no systematic irrationality towards our friends; it merely holds that in some instances, we have reasons of friendship to prevent our (evidentially well-formed) beliefs about our friends from playing their characteristic role in guiding our cognition and action. This is entirely compatible with the view that having a rational and accurate assessment of our friends is in general an essential component of friendship, and it makes no demands that we set out to distort our belief-forming mechanisms. Instead, by accepting something which we cannot on good evidential grounds believe about our friends, we are committing ourselves to a supportive stance towards our friends, in the situations where that is what friendship requires.[67]

Finally, acceptance also allows us to address worries about authenticity. It would indeed seem inauthentic if we systematically believed well about our friends because we thought that was how a friend should believe, rather than out of a systematic attunement to them and their good-making features. However, in the cases of interest, it *is* the very fact that someone is our friend that motivates the doxastic duty—friendship is, in these cases, exactly the kind of reason to which we need to be responsive, despite our evidence. In such cases, responding to the fact of one's friendship does not seem inauthentic; quite the contrary, it seems to be motivated by love for our friends. However, the importance of authenticity also reveals that the cases in which we accept for reasons of friendship need to be the exception rather than the rule—if someone was constantly trying to accept positive things about their friends, we might indeed run back into worries about the self-deception and authenticity of the friendship. As noted above, as a general policy, it is important that we form beliefs about our friends in response to evidence of their good qualities and attributes.

6. What Acceptance Gets Right

I've now argued that acceptance can handle the worries that seemed so pressing when we framed our obligations in terms of believing against the evidence. But there is more work to do to show that acceptance can really capture the rich landscape of cases where we seem to owe our friends

[67] For a similar idea, see Frost-Arnold (2014) on the role of acceptance in proleptic trust. Frost-Arnold draws on Bratman's notion of acceptance in trust where belief is not rationally permitted; I suspect her overall picture would be quite friendly to this cognitive profile of acceptance.

belief against the evidence, and that it can play the roles that care about regarding the stake we have in how our friends think of us.

Our discussion of how acceptance addresses the objections against partialism already began to draw out several positive reasons to think that acceptance accurately captures the landscape in cases of interest. Acceptance allows us to be directly responsive to reasons of friendship as such, it does not depend on a theoretically untenable account of doxastic control, and it makes sense of the observation that the doxastic demands of friendship can vary across contexts. But acceptance goes further towards characterizing the doxastic landscape of friendship by helping clarify two key ideas that partialists have sought to capture: that friendship can make demands not only on our words and actions but also our thoughts,[68] and that in doing so friendship can demand a psychological commitment, and significant cognitive work.

6.1. The Richness of Response Modulation

First, acceptance accounts for the restructuring of diverse behavioral *and psychological* processes. We see this when we consider more precisely just what it means to prevent a belief from playing its default guiding role in cognition and action. Take Mateo's acceptance that Shelby did not cheat on her exam. Some of the consequences of Mateo's acceptance, especially the behavioral manifestations, are relatively straightforward. He might say supportive things about her in conversation with others, or he might still give her an important role in their next group project (and he might do neither of these if he were allowing his belief in her guilt to guide him). He might also prevent himself from reasoning on the basis of her guilt, as he normally would: for example, he needs to prevent himself from concluding that because she cheated her class grade is going to suffer, or that because she cheated on this exam she might be more likely to cheat on the next one as well.

But when we start to consider not only the behavioral effects of acceptance but also the cognitive ones, we need to consider that beliefs normally guide a host of cognitive activities: in addition to guiding reasoning, intention, planning, and action, they also guide our patterns of thought, attention, memory, and other various processes. This means that when someone accepts, they seek to intervene on the way that belief normally guides these varied processes. So not only might Mateo try to prevent himself from making inferences or plans that take Shelby's guilt as a premise—he might also try to prevent himself from spending time thinking or ruminating about her guilt (e.g., suppressing thoughts about "why would she do that?"), prevent his worries about her guilt from affecting his patterns of

[68] Basu (2019b) writes (though her concern is broader than just epistemic partiality): "there is a failure to relate to others as one ought that encompasses not only word and deed, *but also thought* [emphasis added]" (p. 916).

attention (e.g., not letting himself focus on the fact that his texts about her studying progress went unanswered, and instead attending to how determined she was to do better on this test), prevent himself from recalling memories of other times she was guilty (e.g., striving not to dwell on memories of the last time she cheated, and instead recalling other times she's surpassed expectations on tests).

Drawing out these diverse regulatory consequences of acceptance—many of which will feel like familiar parts of trying to be a good friend (and familiar more generally!)—reveals that acceptance accounts for a variety of ways in which we might intuitively think a good friend ought to respond cognitively in these kinds of situations. A friend ought indeed to restructure their patterns of thought, reasoning, and attention—but it turns out that insofar as these processes are normally guided by our beliefs, seeking to restructure these various processes *just is* part of accepting. Acceptance thus unifies what might otherwise appear to be a set of independent cognitive duties of friendship in such cases. Further, the robustness of the psychological regulation drives home that acceptance does not reduce to merely "pretending" or "acting as if" you believe, in a pejorative sense—the goal is not superficially behavioral, or centrally aimed at manipulating the beliefs of others. Rather, it centrally involves a positive goal of maintaining a particular way of thinking about one's friends.

6.2. Acceptance over Time

Acceptance also captures the diachronic profile of partiality cases, in several key ways. First, it captures the profile of commitment and effort. Recall that acceptance is not a one-off action, but rather involves a series of effortful mental control actions exercised over time. Preventing belief's usual role across all these diverse psychological and behavioral patterns is not something one just wills and then it's done completely; it will instead involve continuous maintenance for as long as the underlying belief remains intact and the agent remains committed to regulating its downstream effects. This will take cognitive effort (i.e., engagement of executive control processes) and psychological commitment, especially if the underlying belief is frequently salient and activated.

I think this rightly captures how cases like CHEATING play over time. As long as Mateo still takes his evidence to speak in favor of Shelby's guilt, his trying to "take her side" psychologically will demand a serious kind of psychological commitment.[69] Further, the fact that acceptance is constituted by this extended pattern of effortful mental actions means that there are many opportunities for the accepting agent to *fail* to identify and/or suppress the downstream effects of their belief. There is a non-zero probability of error (whether of monitoring or of suppression) for every cognitive effort

[69] Compare this to the case of emotion regulation: as long as my anger at my friend remains, I will need to continue to work to suppress its effects if I am so motivated.

made to block the downstream effect of belief (see Sripada, 2018, 2021, for further development of this idea in mental action)—and the probability of such error increases when an agent's executive processes are fully engaged or directed elsewhere. It's likely that an accepting agent will at some points fail to prevent the belief from playing its usual role. This, too, seems to reflect how we might imagine the scenario playing out: we can imagine that when Mateo is overwhelmed by other demanding tasks, or distracted, (or even drunk!), he might accidentally slip up and find himself thinking, acting, or saying something that reveals his underlying belief in Shelby's guilt. Overall, on this picture, accepting can be difficult! But rather than being a problem for the view, I propose that this gets things exactly right: when the epistemic chips are down, being a good friend can be hard work.

Two further things fall out of appreciating acceptance as a committed pattern of cognitively effortful actions deployed over time. First, the action profile should be understood as an instance of *self-control*: regulating behavior and cognition in alignment with our practical commitments. Our moral assessment of an accepting agent should be sensitive to this control profile. For instance, given the difficulty of perfect success in acceptance over long periods of time, an agent may not be fully culpable for each of these slips,[70] particularly if they occur when she is engaged in other cognitively demanding tasks that leave fewer resources available for the mental activities demanded by acceptance. And while a friend might still feel upset about such slips, and an agent might still be somehow answerable for them, we ought to recognize that an agent who undertakes a commitment to accepting something on behalf of her friend is doing something morally serious and praiseworthy, even if—given her limitations as a finite cognitive agent—she does not do so perfectly. Further, our *friend* ought to recognize this: a good friend has some obligation to recognize that you are putting in the (cognitive) work for their sake, and that the occasional error is no sign that you're not a true friend.

This also highlights that insofar as acceptance is a commitment to deploying cognitive effort over time, acceptance can be a kind of psychological burden. In this sense, accepting on behalf of our friends shows that one is willing to take up costs as a friend. It requires the use of executive resources that will become unavailable for other effortful tasks. In the domain of emotion, relying on emotion suppression strategies over time can lead to negative psychological consequences, including stress, anxiety, and reductions in wellbeing (e.g., Butler et al., 2003; John & Gross, 2004; Moore et al., 2008); though it has not been empirically tested, we might predict similar effects in the doxastic domain. The costs of acceptance need to be taken into consideration when figuring out how we ought to handle a

[70] See Hendrickx (manuscript) for further development of this idea.

particular situation. But there is nothing mysterious about the claim that friendship can give rise to obligations that are costly to us: indeed, friendship frequently gives us strong reason to do costly things for our friends that we might not owe to just anyone. Acceptance is just another of instance of the ways in which friendship require us to adopt burdens on behalf of our friends.

Second, reflecting on the temporally extended nature of acceptance highlights that the interaction between acceptance and belief is highly dynamic—and that acceptance can, over time, come to influence belief. Restructuring patterns of thought, attention, and action as the *downstream* consequences of a particular belief at a particular point in time, may well cause changes in one's cognitive and doxastic profile at a later point in time, in virtue of the ways they alter how we reason and infer, and what evidence we have and how we think about that evidence. So though it is important, on my view, that the goal of acceptance is not to alter a particular belief state, on pain of being self-undermining, our commitment to acceptance over time may well alter our beliefs—and the line between acceptance and belief may blur over time, especially if patterns of acceptance become learned and habitual.[71]

6.3. Two Worries

Let us now address two possible lingering concerns about the view: one concerning agential coherence of an accepting agent, and another concerning the status of the underlying belief.

My argument contends that an agent can take some body of evidence to support a belief that *p*—but also choose to gate or suppress that belief, preventing it from playing its default guiding role. One might worry whether this puts the agent in an oddly divided state: she concurrently takes herself to have sufficient reason to believe *p* (e.g., that Shelby cheated on the exam), and also to have sufficient reason to act and reason on the basis of not-*p* (e.g., that Shelby is innocent). Does this kind of internal division somehow threaten one's agential coherence?

The parallel case of emotion regulation again helps us here. An agent can (rationally) appraise a situation as warranting an emotion such as anger—but also recognize that expressing that anger outwardly, or even letting it structure her inner reasoning and thoughts—is prudentially or even

[71] Several recent accounts have focused on analyzing the doxastic/epistemic norms of friendship in terms of how we *attend*, rather than how we believe—see Brinkerhoff (forthcoming), Saint-Croix (forthcoming), and Gardiner (manuscript). Brinkerhoff and others primarily discuss attention in its capacity as an upstream mechanism, focused on noticing, inquiring, and evidence gathering. But though what we attend to influences the beliefs we come to have, attention is also importantly guided by our existing beliefs: beliefs guide attention by influencing what we notice, what we seamlessly assimilate, what we are surprised by, and so on. On my view, regulating patterns of attention is *part of* acceptance—but a part of acceptance that may be particularly likely to end up causing changes in the underlying beliefs themselves.

morally inappropriate. This is a deeply familiar situation, and the claim that an agent who seeks to prevent her anger from guiding her reasoning and action as it would if left unchecked, is somehow agentially incoherent lacks bite. Indeed, it is a sign of her status as an integrated agency that she has the capacity to decide whether to let her emotions guide her, or whether she has reason to override them.

The cases of emotion and belief are not so different. Emotions too carry assessments of situations (even beliefs, on some views), and structure and guide our downstream cognition and action in all kinds of ways. And though philosophers and emotion regulation psychologists like to focus on cases where emotional reactions might be distorting, maladaptive, or otherwise undesirable, we very often have strong rational and practical reason to allow our emotions steer us—especially when we have no reason to doubt that our affective systems are well-functioning and properly attuned. Just as we have the capacity, and sometimes reason, to intervene to prevent the guiding role of emotions—and this does not seem agentially suspect—I propose that the same applies in the case of belief. Though guiding reasoning, cognition, and action is a key part of belief's role in our cognitive economies, we are not merely at the whim of our beliefs; accepting is thus a manifestation of our broader agency, rather than a challenge to it.

A second worry concerns doxastic wronging: some philosophers have contended that beliefs *themselves* can be immoral in virtue of their content (Basu 2019b; Basu 2021; Basu & Schroeder, 2019). Are we still failing in some sense as a friend if we don't *actually believe* the accepted propositions?

Consider that on the evidentialist view of belief, what belief really reflects is the state of our evidential situation—it does not reveal all that much about *us*. In contrast, acceptance *does* reveal important things about us: in accepting, we choose to regulate our beliefs in accordance with our practical and moral commitments. Acceptance thus expresses what we value, and allows us to (re)structure our cognitive lives in accordance with those values—it allows us to take a stand on how we want to think and act in the world. Given this, why think that the thin notion of belief—which is primarily a reflection of the evidential situation an agent happens to be in—is what's so morally significant? The decision to accept seems to reflect our agency and character much more richly—and so I remain rather unconvinced that belief *itself* is so important in comparison.

Nevertheless, I am not here arguing against the very possibility of genuine doxastic wronging; one could still hold that there is something morally deficient about believing in such cases, though perhaps through no direct failure of the agent. Still, I suspect that acceptance—perhaps combined

with other kinds of upstream practices—will get very many of us very much of what we want from the moral-epistemic landscape of friendship.

6.4. A Unified Account

A final theoretical advantage of the approach defended here is that acceptance provides a unified framework for reframing diverse problems in the ethics and pragmatics of belief. Acceptance is not special to the case of friendship; it is an independently plausible general-purpose cognitive regulation strategy that we can deploy in any case where we have moral or practical reason not to let belief play its default guiding role. Acceptance may help us make sense of cases where an agent seems to have an evidentially justified prejudiced belief but feels morally unsettled by that belief, or have a role to play in the explanation of scientific communities in which some researchers need to commit to hypotheses that they do not really think are best supported by the evidence. It may help us better understand the landscape in self-deception, ideological beliefs, believing in oneself, or processes of belief change and transition. I take the fact that acceptance might help us in many debates facing a confusing set of considerations—with worries about doxastic control and practical reasons for belief at their heart—to be another consideration in favor of affording it a role the partiality debate.

Taken together, these considerations make a compelling case that in situations where the partialists thinks that friendship seems to demand something of us other than rational belief, it is really *acceptance* that we owe our friends. Acceptance captures the features of these cases that were so puzzling when we tried to cash out partialism in terms of belief, while also explaining the rich cognitive landscape of these kinds of cases.

7. Conclusion

In this paper, we have worked our way through a puzzle plaguing recent literature on friendship and belief: there has been a persistent intuition that we can sometimes owe it to our friends to believe against the evidence, and yet no previous account has been able to directly satisfy this intuition without either relegating everything to upstream inquiry practices, or else running headlong into worries about doxastic control—especially not without compromising the importance of honesty and authenticity in friendship. I have proposed a novel solution to this problem: what we owe our friends in such scenarios is not belief but rather acceptance—blocking the characteristic downstream effects of an unfavorable belief, and committing to reasoning, acting, and thinking on the basis of the accepted proposition. Acceptance allows us to get what we want out of partialism, without falling prey to the objections that plague the belief-focused version of the thesis.

I will close by highlighting two key contributions of this argument. The first is methodological: I have tried to show that appealing to acceptance can help us navigate an otherwise quite confounding debate. I suspect that many tricky debates in the ethics of belief can be fruitfully clarified by incorporating my account of acceptance—especially those topics where worries about doxastic control and responsiveness to non-evidential reasons are a central problem.

The second takeaway concerns friendship itself: I hope to have shown that friendship can demand *significant cognitive work*. It's an essential component of being a good friend that we can commit ourselves to thinking about them in ways that we might not think about others. Though we may not be able to simply will away unwanted beliefs or emotions regarding our friends, we do have the capacity to exercise significant control over the ways in which these (and other) mental states structure our cognitive lives—and so, at least sometimes, being a good friend involves making this effort. On the one hand, the fact that friendship requires this cognitive work is something that few likely want to deny. But on the other hand, really working through this idea reveals something deeply important about the nature of our relationships to others: the needs and good of other people can make significant moral demands on how we structure our cognitive lives. And it is through the ways that we work to structure our cognitive lives that we can reveal our value and care for other people.

Chapter 3

What We Would (but Shouldn't) Do for Those We Love: Universalism versus Partiality in Responding to Others' Moral Transgression[72]

The moral standpoint is characterized by an attitude of impartiality, a refusal to see the life, projects, good, or interests of any particular person (oneself included) as having a greater or lesser value than those of another. (Archard, 1995, p. 129)

[I]t is absurd to suggest that morality requires one to care, or act as if one cares, no more about one's own child than a stranger. (Wolf, 1992, p. 244)

1. Introduction

The quotes above reflect a deep philosophical tension regarding how to make moral decisions involving those closest to us. On the one hand, treating all people equally is a key tenet of philosophical ethical theories. On the other hand, close relationships are deeply important in people's lives, and may produce special moral obligations. These competing considerations can give rise to wrenching moral decisions when loved ones are involved.

For decades, the vast majority of empirical research in moral psychology primarily studied people's judgments regarding anonymous strangers (e.g., Greene et al., 2001, 2009; Haidt, 2001). Yet psychologists now recognize the importance of studying how relationships affect moral cognition, as a growing body of research reveals that people report dramatically different choices in decisions involving friends and loved ones from those involving strangers. One striking context in which this trend emerges is in deciding how to respond to the moral transgressions of others. Weidman et al.

[72] Martha Berg, Susan Gelman, and Ethan Kross are co-authors on the published version of this paper. We thank Chandra Sripada for comments on an earlier version of the manuscript. We also thank the members of the Kross Emotion and Self-Control Lab, the University of Michigan Language and Cognition Lab Group, and University of Michigan Weinberg Institute for Cognitive Science Seminar Series—particularly Rick Lewis—for their feedback on this work. This work was supported by the University of Michigan, Ann Arbor, and the National Science Foundation Graduate Research Fellowship.

(2020) asked participants to imagine witnessing either a close other (e.g., sister, best friend) or a distant other (e.g., dentist, mail carrier) committing a crime. Participants then had to decide whether they would report the perpetrator or lie to protect them when confronted by a police officer. Across several studies, a robust effect emerged: people were much more likely to protect a close (vs. distant) other, and this discrepancy increased with the severity of the crime people observed (also see Waytz et al., 2013).

But do people think it is *morally right* to behave this way? Against the background of Weidman et al. (2020), we take up the currently underexplored question of whether people believe that relationships *should* influence how they respond to others' transgressions. If people think they should treat close and distant others differently, it suggests that people believe moral rules are sensitive to context—an idea that would challenge a common philosophical conception of morality as applying invariantly across people. If, on the other hand, people think they should treat everyone the same regardless of relationship, then this implies a discrepancy between what people think is right and how they would behave. In addition to posing a major challenge to psychological and philosophical theories that rely on consistency, this would demonstrate the striking result that difficult moral decisions about close others may be a domain in which people are particularly likely to fail to do what they think is right. Here we address these questions across four experiments, in the context of decisions about whether to report moral transgressions of a close or distant other.

1.1 The Moral Universalism Hypothesis

One possibility is that people think that they morally should make the same decision regardless of whether the transgressor is a close or distant other (i.e., the Moral Universalism Hypothesis[73]). This idea that moral rules apply equally across people is a widely accepted philosophical principle. All three major philosophical ethical theories—utilitarianism, Kantian deontology, and virtue ethics—incorporate a tenet of impartiality or universalism. Utilitarianism (Mill, 1895) claims that the morally right action is that which maximizes overall happiness, and clearly states that the interests of all people matter equally, regardless of the particular relationships we have with them (see Singer, 1972 for a particularly strong version of this view). Kantian deontology posits that certain acts are always morally impermissible (Kant, 1785). Strict deontological constraints forbid someone from breaking a moral

[73] "Universalism" may not be the best label for this hypothesis. One could hold, after all, that *everyone* ought to show partiality towards close others—meaning that partiality and universalism are consistent. A more appropriate name would, perhaps, be the "Impartiality Hypothesis." However, since a version of this paper is already published using the term "Universalism Hypothesis," I will continue to use that term here.

rule even for someone they love. Finally, Aristotle's twelve basic virtues include *justice*—an important component of which is treating others equally and fairly (Aristotle, 2009, Book V; Kraut, 2018).

Despite universalism's status as the dominant moral framework in philosophical ethics (Archard, 1995; Wolf, 1992), very little empirical work has directly explored whether universalism guides laypeople's moral judgments. However, several lines of research indirectly shed light on the role universalism might play in various kinds of moral decisions. For instance, both adults (Dawes et al., 2007; Fehr & Fischbacher, 2003) and children starting around age eight (Blake & McAuliffe, 2011; Shaw & Olson, 2011) prefer fair treatment for everyone in economic resource allocation games, and will reject unequal offers that advantage themselves or their ingroup (Elenbaas et al., 2016). Yet this research does not speak decisively in favor of universalism, for children also show ingroup and close-relation favoritism when they are in charge of resource distributions (Olson & Spelke, 2008), and often choose to preserve status quo group-based inequalities (Olson et al., 2011).

In further support of universalist considerations, children are sensitive to interpersonal bias in judging contexts. By fourth grade, children (like adults) prefer, and attribute greater fairness to, neutral judges over those who have a personal connection (like friendship) to a contestant, especially in subjective contests (Mills & Keil, 2008). Such work provides evidence for children's and adults' strong commitment to fairness and impartiality across contexts.

Finally, both children and adults are sensitive to a distinction between moral and conventional rules (e.g., Nucci & Turiel, 1978a, 1978b; Rizzo et al., 2018; Smetana, 1981). One feature that distinguishes these categories is that moral rules are judged to apply invariantly to all people across all social contexts, whereas conventional rules are permitted to vary based on the relevant social standards (Smetana, 2013). Although the moral/conventional distinction does not directly address decisions regarding interpersonal relationships, it does speak to the characterization of moral rules as holding universally for all people.

These diverse lines of research shed light on some ways in which universalist principles of equality and impartiality play an important psychological role. However, none directly examines people's commitment to moral universalism in the context of high-stakes moral decisions, as the philosophical ethical theories considered above would prescribe. Although these philosophical theories are normative and not psychologically descriptive, it would be striking if their core universalist tenet were not expressed in folk moral judgments to some degree. Yet this result is far from guaranteed; rather, we will now discuss compelling evidence for a competing hypothesis.

1.2 The Moral Partiality Hypothesis

The alternative possibility is that people believe they *should* protect close others who have committed a moral transgression more than distant others (i.e., the Moral Partiality Hypothesis). In the philosophical literature, defenses of *moral partiality* consist of arguing that people *are* morally justified in treating the people closest to them with special moral concern. This idea is not new. It can be traced back as far as Confucius's account of filial piety, which emphasized the special moral duties owed to family members (Confucius, 2005; Csikszentmihalyi, 2020), and the Ten Commandments, which tell people to "honor thy father and thy mother" (*King James Bible*, 1769/2017).

More recently, philosopher Susan Wolf has commented that it would be "absurd to suggest that morality requires one to care, or act as if one cares, no more about one's own child than a stranger" (Wolf, 1992, p. 244). She motivates this claim with a dramatic example: consider a woman in a boating accident who finds herself nearby two drowning children—one is her own child, and one is a stranger—and the ability to save only one (Wolf, 1992; see also Williams, 1981). Such cases have compelled a number of philosophers to argue that close relationships often give rise to distinctive moral obligations to display preferences towards close others (Archard, 1995; Baron, 1991; Lord, 2016; Scheffler, 2010; Williams, 1981).

The Moral Partiality Hypothesis has more direct empirical support than the Moral Universalism hypothesis. Loyalty is, for instance, a basic dimension in Moral Foundations Theory, which has received extensive support (Haidt & Graham, 2007). More specifically, people judge others who fail to help kin more negatively than those who fail to help strangers (Hughes, 2017; McManus et al., 2020), say that it is more important to help close others (Killen & Turiel, 1998), and believe that people who fail to help those closest to them are less suitable spouses and friends (Everett et al., 2018).

These partialist tendencies emerge early: developmental researchers have found that by age eight, children judge someone who fails to help their friend (vs. stranger) to be meaner, suggesting that they believe an unhelpful friend is failing their obligations in a way that an unhelpful stranger is not. Conversely, children judge someone who helps a stranger (vs. friend) to be nicer, suggesting that helping a friend is expected, but helping a stranger surpasses one's obligations and is thus especially commendable (Marshall et al., 2020).

We thus also have promising empirical and philosophical support for the partialist idea that people will say they should protect close others more than distant others. Yet the studies discussed here focus primarily on evaluations of others' helping behavior. Our primary question, in contrast, is how people weigh partialist and universalist considerations in their *own* moral decisions regarding close

others (and distant) who have committed transgressions—a question that has not yet been explored in the literature.

1.3 Consistency across Judgments

Thus far we have discussed competing hypotheses regarding whether people will judge that it is right to show moral preference for close others. However, the present research question of whether people think they *should* protect close others who have committed a crime arises not in isolation, but against the backdrop of Weidman et al. (2020)'s findings that people think they *would* protect close others more than distant others. Thus, there is an additional remaining dimension not captured by the discussion thus far: *are people's judgments about what they think they would do consistent with what they think is morally right?*

Much research has established that people are motivated towards maintaining consistency (Festinger, 1957; Higgins, 1987), and a consistently positive moral self-concept (Dunning, 2007; Jordan et al., 2011; Mazar et al., 2008; Monin & Jordan, 2009). Yet admitting one would not do what one believes is morally right would likely challenge one's moral self-concept; thus, people may be psychologically motivated to avoid such an admission. One possible strategy for avoiding this psychological discomfort would be to bring these judgments into alignment: to change one's prediction about how one would behave to match what they judge to be morally right, or to convince themselves that what they would do is the morally right choice. Thus, in support of these psychological goals, we might expect people to report little difference between what they would and should do—especially if they are asked to make both judgments concurrently. Given Weidman et al. (2020)'s findings that people would protect close others more than distant others, this would imply that people also would say that they *should* protect close others more. Accordingly, predicting consistency across "should" and "would" judgments implies either predicting Moral Partiality, or predicting that people's universalist judgments would lead them to change their predictions about how they would act towards close vs. distant others, to bring their "should" judgments into conformity with their "would judgments."

In contrast, if the Moral Universalism Hypothesis is supported and we see the same pattern of behavioral predictions as Weidman et al. (2020), this would leave us with a notable discrepancy between what people think is right, and how they would act. Past research has found that judgments about moral norms, and choices about how to act, arise from distinct psychological processes, and factors that influence one kind of judgment may have little effect on the other (Pletti et al., 2017; Sood & Forehand, 2005; Tassy, Deruelle, et al., 2013; Tassy et al., 2012; Yu et al., 2018). Accordingly, it is

reasonable to predict that relational closeness might impact these distinct judgments in different ways. There is even some initial evidence supporting this possibility: Kurzban et al. (2012) found that in trolley problems, more people say both that they would sacrifice one person to save five others *and also* that doing so is wrong, when the people at stake were kin or friends, than when they were strangers (also see Tassy, Oullier, et al., 2013).

Thus, discrepancies across what people should and would do in response to others' moral transgressions would not be without empirical precedent. Such discrepancies suggest that moral decisions about close others may be a domain in which people are particularly likely to fail to do what they think is right. Given how ubiquitous moral decisions involving close others are in real life, this finding would be noteworthy and indicate an important avenue for future investigation, especially insofar as we think it is important for people to successfully live up to their own moral standards.

1.4. The Present Studies

Across four studies, we seek to adjudicate between the Moral Universalism and Moral Partiality Hypotheses, in the context of responding to others' moral transgressions. As Weidman et al. (2020) argue, witnessing a close other transgress presents people with a unique dilemma between considerations of loyalty and a desire to protect those closest to us and considerations of justice and punishing immoral acts. In contrast, when a distant other transgresses, there is no such conflict. Weidman et al. (2020)'s findings show that when predicting how they *would* behave if faced with this dilemma, loyalty drives people's decisions. In the present work we test whether this pattern persists when people consider what the morally right response is. Thus, our driving question is: do people think that they *should* lie to the police to protect close others from punishment for a crime, in addition to thinking that they *would* do so?

We examined this question by first asking participants to make just one kind of judgment across a series of dilemmas (Studies 1a and 1b), and then asking them to make both judgments about each dilemma (Study 2). In our final study, we undertook a more in-depth examination of participants' "should" judgments (Study 3). All studies were preregistered on As Predicted, and all preregistered methods were followed unless noted otherwise in the text. All studies were determined to be exempt by the University of Michigan IRB; informed consent was obtained from all participants prior to survey administration.

2. Study 1a

We started with the most basic version of our question: do participants who are asked what they *would* do in response to others' moral transgressions give different responses than those asked

what they *should* do? Participants were asked to imagine a series of scenarios in which a close (or distant) other committed a high-severity moral transgression, and were asked either what they would, or should, report the transgression. We expected to replicate Weidman et al. (2020)'s findings that people believe they *would* protect close others more than distant others. Our competing hypotheses deliver distinct predictions for "should" judgments: the Moral Partiality Hypothesis predicts that people believe they should protect close others more than distant others. In contrast, the Moral Universalism Hypothesis predicts that people believe they should protect close and distant others equally.

This study was preregistered through As Predicted #31495 (anonymous link for peer review: https://aspredicted.org/blind.php?x=ab7ka2).

2.1. Method

2.1.1. Participants

Four hundred and three English-speakers in the United States were recruited through Amazon Mechanical Turk (61% women, 39% men M_{age} = 37.24, SD = 12.31). The self-reported racial/ethnic breakdown of the sample (participants could select multiple categories) was: 7% Asian, 9% Black or African American, 5% Hispanic or Latino, 2% Native American, 79% white, and 1% other.

Participants were excluded from analysis according to the following criteria: answering "no" to a validity check question (N = 4); saying that English was not their native language (N = 4); providing the same name for multiple close/distant other nominations (N = 16); and failing the manipulation check (N = 80), for a final sample of 299. Due to oversight, manipulation check failure was not included as a pre-registered exclusion criterion; regardless, all results presented below are statistically equivalent when we include participants who failed the manipulation check.

2.1.2. Procedure

Participants were randomly assigned to either the "would" or "should" condition. They were first asked to think of two people whom they considered the closest to them (e.g., father, spouse, sister, best friend) and two of their most distant acquaintances (e.g., mailman, landlord, dentist). For each, participants provided a first name and the nature of the relationship.

Next, participants were presented with eight vignettes. In each, they were asked to imagine that they had witnessed one of the nominees committing a high-severity theft, such as stealing a laptop or a wallet (see Appendix A for full list of scenarios). Participants were then asked to imagine a police officer approaching them and asking whether they had seen anything suspicious. (Henceforth, we will call these "punish-or-protect" dilemmas, for convenience.) Depending on their assigned condition,

they were asked either whether they *would* report the transgressor ("what you really would do"), or whether they *should* report them ("the ideal, right thing to do"). Participants responded using a 6-point Likert scale (1 = "Definitely would/should not report"; 6 = "Definitely would/should report"). After the final trial, they were asked to write about their thought process as they decided how to answer the preceding question; open-ended data was collected for exploratory purposes and was not analyzed for the present research.

A manipulation check asked participants which kind of judgment they had been asked to make. Participants then reported their level of trust in the police (0 = "Very untrustworthy"; 6 = "Very Trustworthy") as an exploratory measure, and completed a set of demographic questions.

2.2. Results

To test the effects of judgment and relationship on reporting, we ran a mixed linear model using the lme4 package in R (Bates et al., 2015). We included participant and dilemma as random intercepts, which was the maximal model that reached convergence. Likelihood of reporting the act was reverse-coded, so that higher scores indicate a greater likelihood of protecting the perpetrator. We reverse-code reporting as protecting in all studies presented here for the sake of consistency with Weidman et al. (2020)'s methods, and to facilitate comparison across these studies. We probed the interaction between relationship and judgment type with four follow-up tests, using a Tukey correction for multiple comparisons (corrected *p*-values reported).

Overall, participants' responses indicated more protection for close others (M = 3.29, SD = 1.85) than distant others (M = 2.01, SD = 1.35; b = 1.23, $t(2084.74) = 27.78$, $p < .001$, 95% CI = [1.15, 1.31]). This pattern emerged for both "would" judgments (simple effect: b = -1.49, $t(2084.66) = 26.19$, $p < .001$, 95% CI = [-1.71, -1.27]) and "should" judgments (simple effect: b = -0.97, $t(2085.15) = 14.31$, $p < .001$, 95% CI = [-1.23, -0.71]), supporting the Moral Partiality Hypothesis.

However, people also reported that they *would* protect a transgressor (M = 2.83, SD = 1.82) more than they reported that they *should* (M = 2.39, SD = 1.59; b = -0.44, $p < .01$, 95% CI = [-0.74, -0.14]). This difference was greater when the transgressor was a close other (simple effect: b = 0.70, $t(352.67) = -4.53$, $p < .01$, 95% CI = [0.15, 1.25]) versus a distant other (where the difference was not significant; simple effect: b = 0.18, $t(352.67) = -1.84$, $p = .24$, 95% CI = [-0.37, 0.18]; closeness x judgment interaction: b = -0.52, $t(2085.16) = -5.82$, $p < .001$, 95% CI = [-0.70, -0.34]). These results suggest that there is a discrepancy between what people believe they would do and what they think is morally right regarding close others—indicating that while there is partiality in "should" judgments, that partiality is weaker than for "would" judgments.

All effects of relationship and judgment were equivalent in an exploratory model that added police trust (M = 3.96, SD = 1.50 on a 0-6 scale) as a covariate; full model statistics are reported in a Supplement in Appendix C.

3. Study 1b

Study 1b served two purposes. First, we sought to replicate the findings of Study 1a. Second, we added an explicit contrast between "would" and "should" judgments in the instructions given to the participants. Whereas in 1a, participants had read a description of only the one type of judgment they were asked to make (either "would" or "should" judgments), in 1b, participants read about both types before learning which kind of judgment they were to make. This further clarified the instructions and made an explicit contrast between the judgment types.

This study was preregistered through As Predicted #31493 (anonymous link for peer review: https://aspredicted.org/blind.php?x=gr3ev8).

3.1. Method

3.1.1. Participants

Three hundred and ninety-nine native English-speakers in the United States were recruited through Amazon Mechanical Turk (56% women, 43% men, 0.3% other; M_{age} = 39.15, SD_{age} = 12.79). The self-reported racial/ethnic breakdown of participants (where participants could select more than one option) was: 7% Asian, 7% Black or African American, 6% Hispanic or Latino, 1% Native American, 0.3% Native Hawaiian or Pacific Islander, 78% white, and 2% other.

Participants were excluded from analysis according to the same criteria as in 1a: non-native English speakers (N = 5); providing the same name twice (N = 19), and failing the manipulation check (N = 59). This gave us a final sample of N = 316. Results did not differ when we included participants who failed the manipulation check.

3.1.2. Procedure

The design of Study 1b was nearly identical to Study 1a, again using a 2 (relationship: close, distant; within-subjects) x 2 (judgment: should, would; between-subjects) design. The only difference was in the instructions: while in 1a, participants had simply been asked to make either "would" or "should" judgments, in 1b the instructions explicitly contrasted the two kinds of judgments. The instructions read, "In such situations, people may think about what they should do (the ideal, right thing to do) or they may think about what they would actually do (how they would behave in the real world)," and then told participants which kind of judgment they would be asked to make.

3.2. Results

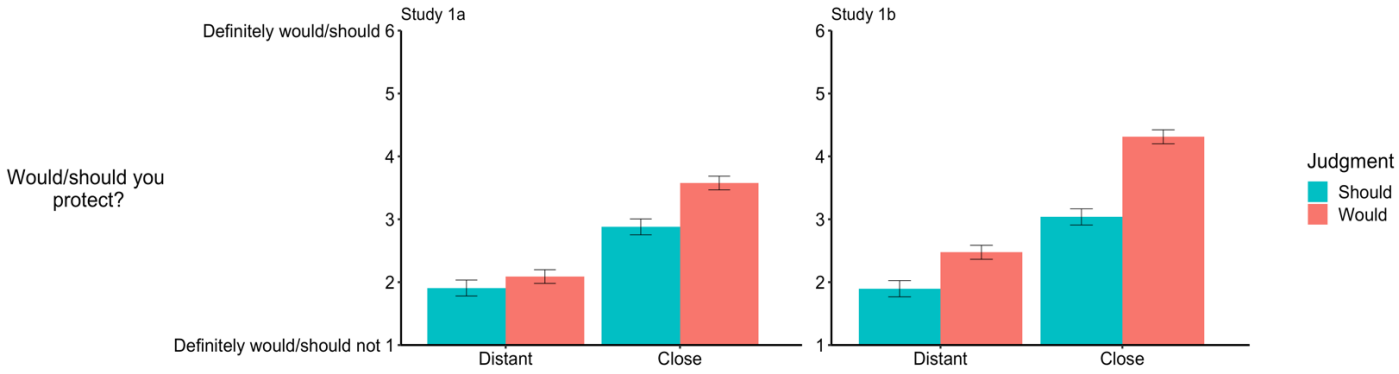
We used the same mixed linear model as in 1a, again including participant and dilemma as random intercepts, which was the maximal model that reached convergence. Likelihood of reporting was reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. We probed the interaction between relationship and judgment with four follow-up tests, using a Tukey correction for multiple comparisons (corrected *p*-values reported).

Overall, we replicated our results from 1a. Participants were significantly more inclined to protect close others (M = 3.78, SD = 1.94) than distant others (M = 2.23, SD = 1.57; b = 1.49, $t(2203.21) = 33.10$, $p < .001$, 95% CI = [1.41, 1.57]). This pattern again emerged for both "would" judgments (simple effect: b = -1.84, $t(2205.49) = 31.54$, $p < .001$, 95% CI = [-2.06, -1.62]) and "should" judgments (simple effect: b = -1.14, $t(22.04.52) = 16.62$, $p < .001$, 95% CI = [-1.40, -0.88]).

As in 1a, people also revealed that they *would* protect a transgressor (M = 3.39, SD = 1.98) more than they *should* (M = 2.47, SD = 1.72; b = -0.93, $t(314.00) = -6.07$, $p < .001$, 95% CI = [-1.23, -0.63]). The difference was again greater when the transgressor was a close other (simple effect: b = 1.27, $t(370.68) = -8.00$, $p < .001$, 95% CI = [0.69, 1.85]), versus a distant other others (simple effect: b = 0.58, $t(370.68) = -3.64$, $p = .05$, 95% CI = [-0.004, 1.16]; relationship x judgment interaction: b = -0.69, $t(2206.41) = -7.71$, $p < .001$, 95% CI = [-0.87, -0.51]; see Figure 3.1).

Figure 3.1

Effect of Relationship on "Would" and "Should" Judgment: Studies 1a and 1b



Note. Error bars show standard error of the mean.

Results were equivalent when we controlled for police trust (M = 3.83; SD = 1.58); see Supplement in Appendix C for full model.

Finally, we note that across both Studies 1a and 1b we observed that more people failed the manipulation check—which asked them what kind of judgment they had been asked to make—in the "should" condition (after other exclusions: 89% of failures in 1a; 85% of failures in 1b). Of those

"should" failures, most said that they had been asked to say what they *would* do (96% in 1a; 92% in 1b). We interpret this pattern as resulting from the fact that "what should you do" is sometimes used in everyday language to ask a predictive question (about how one is likely to act) rather than a strictly normative one (about how one ought to act). This ambiguity between readings is absent for "would," which asks a clearly predictive question. Thus, although our opening instructions asked people what they *ideally should do*—a normative, deontic question—some participants may have strayed towards the predictive reading as they completed the survey. Excluding these participants did not alter the results.

3.3. Studies 1a and 1b Discussion

In Studies 1a and 1b, we tested whether participants who were asked what they would do in response to a close (vs. distant) others' moral transgression gave different responses than those who were asked what they should do. In both studies, we replicated Weidman et al. (2020)'s findings that participants would protect close others more than distant others. Participants also said they *should* protect close others more – providing initial support for the Moral Partiality Hypothesis.

However, we also saw a notable discrepancy emerge across judgments: people thought they *would* protect close others more than they *should*, due to the fact that relationship affected "would" judgments more strongly than "should" judgments. These findings thus suggest that people are relatively less partialist when thinking about what is morally right, as opposed to how they would actually act. This resulting discrepancy between judgments also provides initial evidence that moral decisions involving close others may indeed be a context in which people are especially unlikely to do what they think is right.

4. Study 2

Studies 1a and 1b provide initial evidence for two important claims. First, that people believe they should protect close others more than distant others. And second, that relationships influence what people think they should do more weakly than what they think they would do. These findings suggest that while people believe that relationships influence what the morally right decision is, there is also discrepancy between what people think is right and what they would actually do when it comes to close others.

In Study 2, we sought to test the strength of this discrepancy between "would" and "should" judgments. In Studies 1a and 1b, participants made only one type of judgment, meaning there may not have been much psychological pressure for people to avoid a discrepancy between what they would do and what they say is right. If participants were asked to make both "would" and "should"

judgments together, it is possible that they would feel an increased pressure to make those judgments consistent, in order to avoid admitting that they would fail to act morally (by their own lights).

Participants in Study 2 were thus asked both what they would and should do for every dilemma. (See Appendix E for two supplemental studies in which participants made only one judgment about each dilemma, but where judgment type is varied across dilemma within subjects.) For the discrepancy between judgments to emerge in this within-subjects design, participants would have to be both *self-aware* of this discrepancy, and also *willing to admit* that they would not do what they think is morally right.

Based on our previous results, we predicted that we would still see higher protection of close than distant others, for both judgment types. We further predicted that relationship would influence "would" judgments more strongly than "should" judgments, leading to a discrepancy between what people think they would and should do regarding close others.

This study was preregistered through As Predicted #38203 (anonymous link for peer review: https://aspredicted.org/blind.php?x=xe76ve).

4.1. Method

4.1.1. Participants

Four hundred and two native English-speakers in the United States were recruited through Amazon Mechanical Turk (51% women, 48% men, 0.7% other; M_{age} = 36.38, SD_{age} = 11.58). The self-reported racial/ethnic breakdown of participants (where they could select more than one option) was: 7% Asian, 11% Black or African American, 4% Hispanic or Latino, 0.5% Native American, 77% white, and 1% other.

Participants were excluded for being non-native English speakers (N = 8), responding that their data was not valid (N = 3), and giving the same name more than once (N = 35; all pre-registered exclusion criteria); for a final sample of N = 356.

4.1.2. Procedure

Participants again nominated two close and two distant others, and considered the same eight punish-or-protect dilemmas as in the previous studies. Instructions included the explicit contrast between "should" and "would" judgments. After each dilemma, participants were asked both what they should and what they would do. Both questions were presented on the same screen; question order was held constant across dilemmas for each participant, but counterbalanced across participants.

4.2. Results

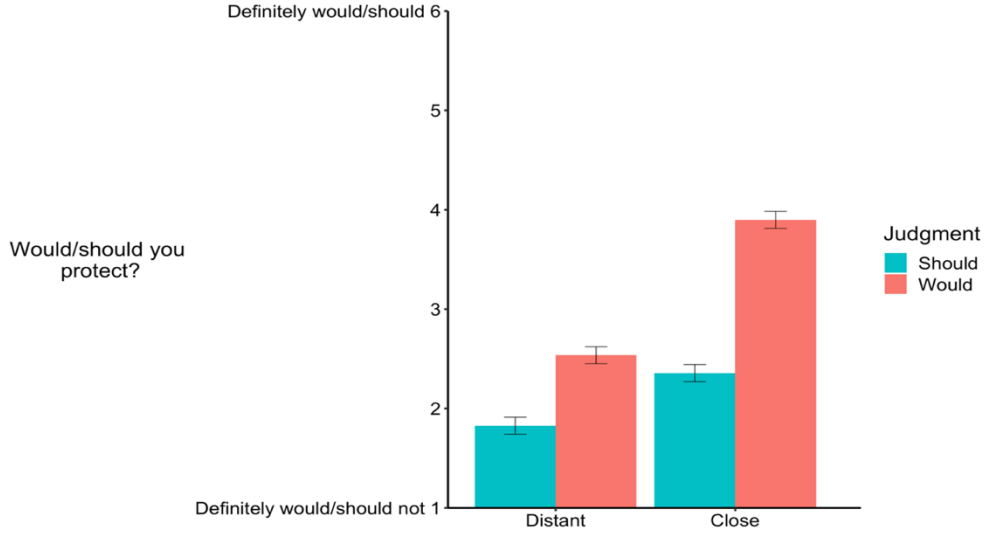
We used the same linear mixed model method as in previous studies, including participant and dilemma as random intercepts as the maximal model that reached convergence. Likelihood of reporting was again reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. There was no significant effect of question order across participants; accordingly, we collapsed across order for the analyses reported here. We probed the interaction between relationship and judgment with four follow-up tests using a Tukey correction for multiple comparisons (corrected *p*-values reported).

We saw largely the same pattern of results as in Studies 1a and 1b. Again, participants showed higher protection for close others (M = 3.13, SD = 1.93) compared to distant others (M = 2.18, SD = 1.55; b = 0.95, t(5328.38) = 30.75, p < .001, 95% CI = [0.89, 1.01]). This pattern emerged for both "would" judgments (simple effect: b = -1.36, t(5328.190) = 31.31, p < .001, 95% CI = [-1.51, -1.21]) and "should" judgments (simple effect: b = -0.53, t(5328.19) = 12.17, p < .001, 95% CI = [-0.68, -0.38]), further supporting the Moral Partiality Hypothesis.

We again saw a difference between what people thought they would do (M = 3.22, SD = 1.89) and should do (M = 2.09, SD = 1.54; b = -1.13, t(5327.99) = -36.62, p < .001, 95% CI = [-1.19, -1.07]). This difference emerged for both close (simple effect: b = 1.54, t(5327.99) = -35.47, p < .001, 95% CI = [1.39, 1.69]) and distant others (simple effect: b = 0.71, t(5327.99) = -16.32, p < .001, 95% CI = [0.56, 0.86]), but was greater for close others (interaction: b = -0.83, t(5327.99) = -13.54, p < .001, 95% CI = [-0.95, -0.71]; see Figure 3.2).

Figure 3.2

Study 2: Would/Should Judgments in a Within-Subjects Design



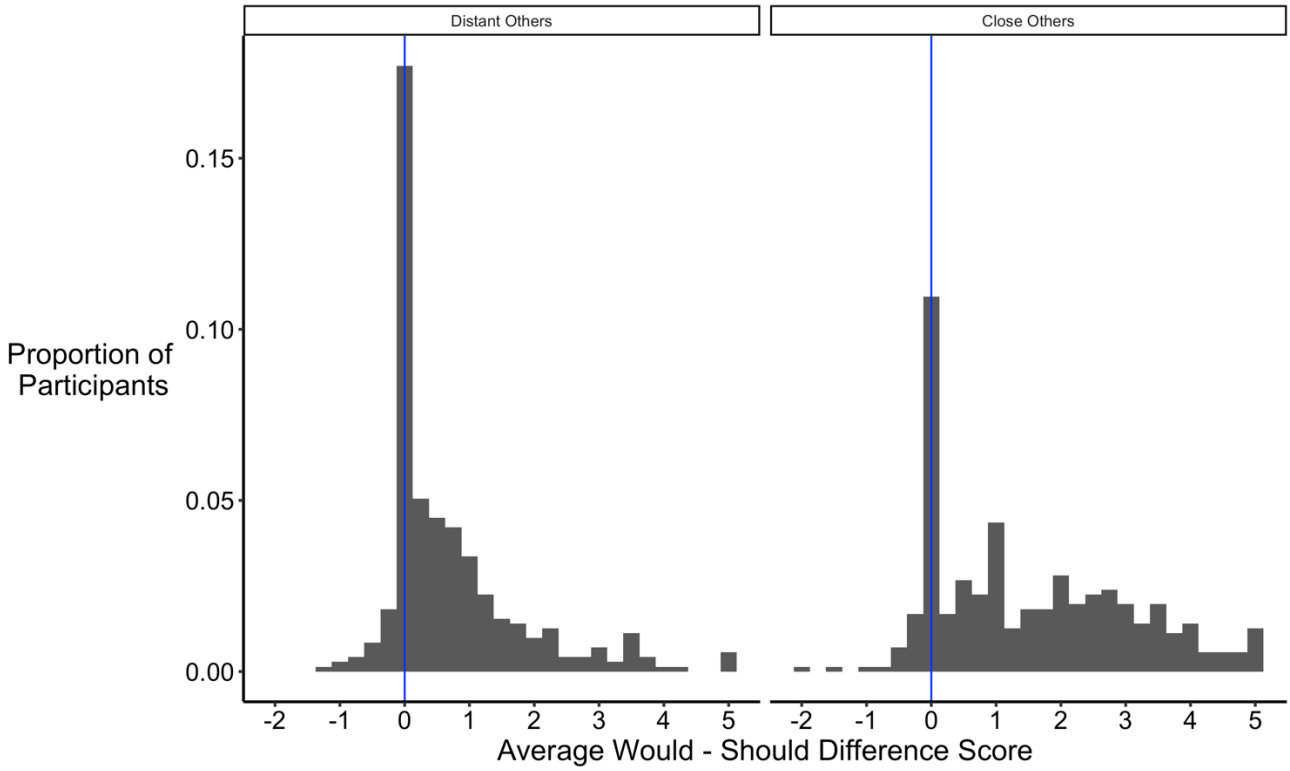
Note. Error bars show standard error of the mean.

Results were equivalent when we controlled for police trust (M = 3.91, SD = 1.61); full model statistics reported in a Supplement in Appendix C.

To further explore the distribution of would/should discrepancies, we calculated each participant's average would-should difference for close and distant others; this distribution is plotted in Figure 3.3. This visualization reveals a group of participants who showed no difference between what they reportedly would and should do (though, consistent with our primary analyses, a far larger number of participants showed no average would/should discrepancy for trials involving distant others), and a group who indicated a difference, though to varying degree. Future research might explore the psychological predictors and effects of these individual differences.

Figure 3.3

Would-Should Difference Scores for Close vs. Distant Others



Note. Difference scores by participant, calculated by subtracting "should" judgments from "would" judgments for each vignette and averaging across vignettes, separately for close vs. distant others.

4.3. Discussion

Study 2 revealed largely the same pattern of results as Studies 1a and 1b. In further support of the Moral Partiality Hypothesis, participants again said that they both would, *and should*, protect close

others more than distant others. Relationship again influenced "should" judgments more weakly than "would" judgments, resulting in the key finding of Study 2: that the discrepancy between what participants said they would and should do persisted even when they were asked to make both judgments about each dilemma. Given what we know about people's motivation to maintain a positive and consistent moral self-concept, this discrepancy across judgments is striking: it suggests that people recognize—and are willing to admit—that they would not do what they think is right by their own lights. This finding further supports the hypothesis that difficult moral decisions about close others may be a domain in which people are particularly likely to fail to act as (they believe) they should.

5. Study 3

Thus far, our evidence supports the Moral Partiality Hypothesis: people believe that they should be more lenient regarding close others' moral transgressions. However, we have also seen that relationship influences what people think is right *less* than it influences what they think they would do. This finding suggests delivers a key upshot: participants have admitted to a striking *discrepancy* between what they say they would and should do, especially when it comes to close others.

There remain, however, two available—and importantly different—interpretations of this discrepancy, both of which are consistent with the evidence thus far. The first is that people genuinely think that they *would not* do what they think they *should do*; that is, they are revealing an inconsistency between their predicted actions and their evaluative judgments. Natural extensions of such a finding would include discussions of whether, in displaying this inconsistency, people are being irrational, hypocritical, or akratic (acting against their own considered best judgments).

An alternative view appeals to a "Many Reasons" picture of judgment, which is a widespread (though not entirely uncontroversial) view in philosophy. On this view, one's overall set of reasons to rationally choose to do something include moral reasons (e.g., whether it would harm someone) as well as non-moral "pragmatic" reasons (e.g., whether it would advance one's own personal goals). It follows that a person can rationally and consistently choose to do something even if the moral reasons oppose it, because they are outweighed by pragmatic reasons. That is, all things considered, they "overall" should do that thing, even though they *morally* should not.

To give a concrete example, consider someone at the airport who notices a passport on the ground. Under normal circumstances, they probably ought (morally and overall) to turn it in to security so it can be safely returned to its owner. But on this day, doing so would come at excessive cost: if they take the time to turn in the passport, they will miss their flight. So while they still might have moral reason to turn the wallet in, it is reasonable to say—taking all the relevant reasons into account

(moral and pragmatic) —that they ought to leave it and go catch their flight. If, like the passport-finder, people distinguish what they *morally* should do from what they *overall* should do, then perhaps there is no inconsistency at all in our observed difference between "would" and "should" judgments. In some circumstances, it may be that failing to act on our moral reasons over practical ones is selfish or otherwise reveals bad character, but there may be nothing *inconsistent* or *irrational* about it.

In Study 3, we sought to adjudicate between these two possible interpretations. We also addressed a related question arising from wording of the "should" questions used thus far: asking people about the "ideal, right thing to do" is arguably ambiguous between the "overall should" and "morally should" interpretations. Accordingly, in this study we asked participants to make one of four judgments about the set of punish-or-protect dilemmas: participants were either asked about what they *actually would* do, *ideally should* do (the language used throughout Studies 1a, 1b, and 2), *morally should* do, or *overall should* do.

We propose two competing hypotheses for this study, based on the conceptual possibilities outlined above. The Genuine Inconsistency Hypothesis predicts that participants will say they *would* protect close others more than they morally *and overall* should protect them. In contrast, the Many Reasons Hypothesis predicts no discrepancy between what people think they would and overall should do, instead revealing that people say they *overall should* protect close others more than they *morally should*.

Finally, it is possible that we will see something in between these two clear hypothesized alternatives: that "overall should" judgments will fall somewhere between "would" and "morally should" judgments. This would suggest that people can distinguish between "should" judgments that encompass a broad range of reasons and those that include merely moral ones, but also that there is an inconsistency between what people think they ought to do and actually would do. This study will serve the additional purpose of revealing whether people were interpreting our "ideally should" question in the prior studies as a specifically moral question (as we intended), or whether they were taking it to ask about a broader, all-things-considered "should" judgment.

This study was preregistered through As Predicted #43538 (anonymous link for peer review: https://aspredicted.org/blind.php?x=qq6tt5).

5.1. Method

5.1.1. Participants

Seven hundred and ninety-eight native English-speakers from the United States were recruited through Prolific (51% women, 47% men, 2% other; M_{ags} = 33.34, SD_{ags} = 11.91). This sample size entailed 200 participants per cell, before exclusions, as in all previous studies. The self-reported

racial/ethnic makeup of participants was: 9% Asian, 8% Black or African American, 6% Hispanic or Latino, 0.4% Native American, 0.4% Native Hawaiian or Pacific Islander, 72% white, and 2% other.

Participants were excluded from analysis according to the following criteria: duplicate worker IP addresses (cutting just the second response, $N = 4$); non-native English speakers ($N = 3$); not responding to all dilemmas ($N = 1$); saying their data were not valid ($N = 9$); providing the same name more than once ($N = 15$); gibberish or nonsensical answers to the open-ended study probe ($N = 10$); being in the wrong "ballpark" for the judgment manipulation check (e.g., selecting "would" when they were in one of the "should" conditions; $N = 147$).[74] This gave us a final sample of $N = 609$.

5.1.2. Procedure

The basic paradigm was the same as in all previous studies, but with two additional levels in the judgment factor, giving us a 2 (relationship: close, distant; within-subjects) x 4 (judgment: ideally should, morally should, overall should, would; between-subjects) design. Before the dilemmas, participants were presented with one of four sets of instructions telling them what kind of judgment to make: what you ideally should do; what you morally should do; what you overall should do; or what you actually would do; see Table 3.1 for full instructions. Because this study depended on participants clearly understanding which judgment they were supposed to be making, additional language was included in the presentation of each dilemma reminding participants of which judgment they were being asked for. After participants responded to the eight dilemmas, they completed an exploratory open-ended response question, an exploratory police trust measure, a judgment manipulation check, an attention check, demographic questions, and an open-ended study probe.

[74]We preregistered that we would exclude participants who failed an attention check question, which was disguised as an additional police trust question but asked participants to type a particular response. However, only 304 participants responded correctly; it is possible that it was "too well hidden" as a closing demographic question that came after all the main experimental survey questions were completed. Accordingly, we did not use this question as an exclusion criterion for the analyses reported here.

Table 3.1

Instructions for Each Judgment

Judgment	Instructions
Would	We will be asking you to think about what you **actually would do** – how you would behave in the real world, if you really found yourself in this situation.
Ideally Should	We will be asking you to think about what you **ideally should do** – the ideal, right thing to do.
Morally Should	We will be asking you to think about what you **morally should do –** what the most morally right choice is in this situation.
Overall Should	We will be asking you to think about what you **overall should do** – what the all-things-considered best decision is, when you account for all the factors and complexities of the decision, including both moral and practical considerations.

5.2. Results

Likelihood of reporting was again reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. To test our hypotheses, we fit a multilevel model predicting protecting from the type of judgment people made, which was coded as a set of three orthogonal contrasts to test our three primary research questions: do "would" judgments differ from all kinds of "should judgments; do "overall should" judgments differ from "morally should" and "ideally should" judgments; and do "ideally should" and "morally should" judgments differ from each other. The maximal model that reached convergence included participant and dilemma as random intercepts, and relationship as a random slope across participants. To interpret interactions, we ran four follow-up simple effects tests, using a Tukey correction for multiple comparisons (corrected *p*-values reported). Descriptive statistics for cells are reported in Appendix B.

We replicated the main effect of relationship across judgment types, with overall higher protecting for close others (M = 3.04, SD = 1.82) than distant others (M = 2.01, SD = 1.42; b =1.01 $t(604.79) = 19.50$, $p < .001$, 95% CI = [0.91, 1.11]).

The first contrast tested whether "actually would" judgments differed from the three kinds of "should" judgments. "Would" judgments (M = 3.17, SD = 1.84) elicited higher protecting responses than all "should" judgments (M = 2.26, SD = 1.58; b = 0.89, $t(605.01) = 7.97$, $p < .001$, 95% CI = [0.67, 1.11]), consistent with the Genuine Inconsistency Hypothesis. As in previous studies, the discrepancy between what people said they would and should do was greater for close others (simple effect: b = 1.20, $t(702.85) = 10.37$, $p < .001$, 95% CI = [0.96, 1.44]) than distant others (simple effect:

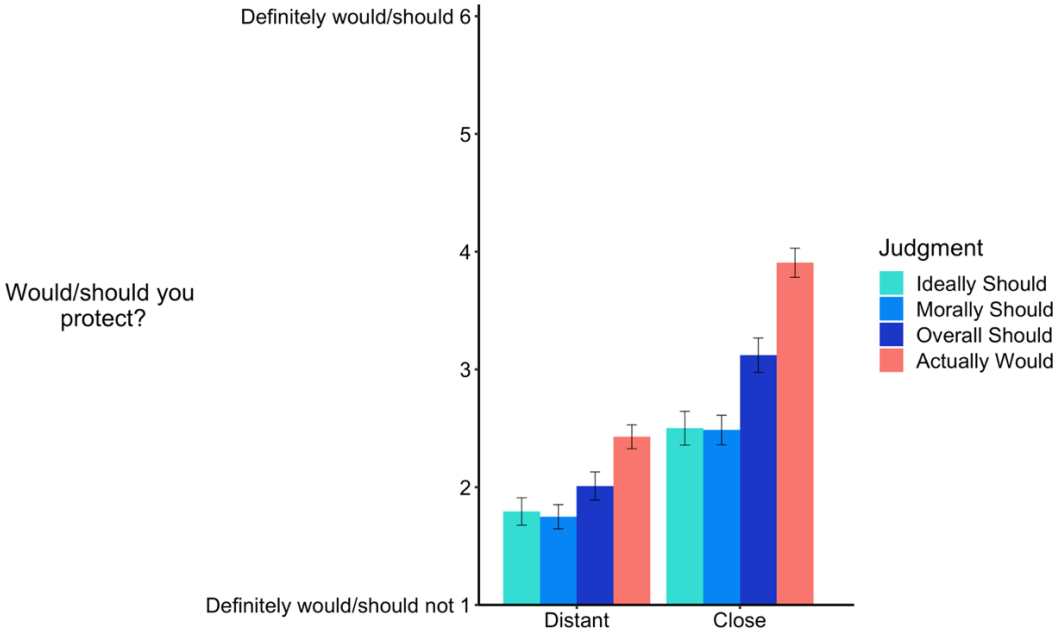
b = 0.56, t(702.86) = 4.98, p < .05 , 95% CI = [0.32, 0.80]; contrast 1 x relationship interaction: b = 0.63, t(605.29) = 5.60, p < .001, 95% CI = [0.41, 0.85]). (See Appendix D for a Supplement containing an exploratory analysis comparing only "overall should" and "would" judgments.)

The second contrast tested whether participants' judgments of what they *overall should* do differed from judgments of what they *morally should* do. "Overall should" judgments (M = 2.57, SD = 1.65) elicited higher protecting responses than "morally should" and "ideally should" judgments (M = 2.13, SD = 1.53; b = 0.43, t(605.01) = 3.23, 95% CI = [0.18, 0.68], p < .01). The difference between what people overall and morally/ideally should do was greater for decisions about close others (simple effect: b = 0.63, t(702.65) = 1.72, 95% CI = [0.36, 0.90], p < .01) than distant others (simple effect: b = 0.24, t(702.65) = 1.72, p = .93, 95% CI = [-0.03, 0.51]; contrast 2 x relationship interaction: b = 0.39, t(604.78) = 2.90, p < .01, 95% CI = [0.14, 0.64]).

Finally, the third contrast tested for differences between "morally should" (M = 2.12, SD = 1.57) and "ideally should" judgments (M = 2.15, SD = 1.48). No significant difference emerged (b = -0.03, t(605.01) = -0.21, p = .84, 95% CI = [-0.32, 0.26]). Results for Study 3 are shown in Figure 3.4.

Figure 3.4

Study 3: Protecting Decisions across Four Judgments



Note. Error bars show standard error of the mean.

Results were equivalent when we controlled for police trust (M = 3.07, SD = 1.78); full model statistics reported in the Supplement in Appendix C. Results were also equivalent when we applied a

more stringent manipulation check exclusion criterion, excluding any participant who truly failed the manipulation check, rather than just those who were in the wrong "ballpark" category (N = 530).

5.3. Study 3 Discussion

Study 3 examined whether the discrepancy between "should" and "would" judgments documented in Studies 1a, 1b, and 2 reflected a genuine inconsistency between what people thought they should and would do in punish-or-protect dilemmas (the Genuine Inconsistency Hypothesis), or whether participants were *not* showing any inconsistency, but instead thought that what they morally should do is different from what they all-things-considered should do (the Many Reasons Hypothesis).

Consistent with the Genuine Inconsistency Hypothesis, "would" judgments elicited higher protecting responses than all three kinds of "should" judgments, especially for close others. Additionally, "overall should" judgments elicited higher protecting than "morally should" and "ideally should" judgments, particularly for close others. This suggests that people distinguish between specifically moral "oughts" and all-things-considered normative judgments. The fact that people said they *would* protect more than they overall *should* suggests that there is a genuine inconsistency across people's judgments—and that this inconsistency cannot merely be accounted for by appealing to a broader kind of "should" judgment.

Finally, our analyses revealed no difference between "morally should" and "ideally should" judgments, suggesting that the judgments participants have been making across all our studies have indeed been judgments about what they morally should do.

6. General Discussion

Across four studies, we examined whether people believe they not only would, but also *should*, preferentially protect close others who have committed severe moral transgressions. In doing so, we tested two competing hypotheses—the Moral Universalism and Moral Partiality Hypotheses—each of which carries an important, but distinct, insight into the nature of moral decision-making.

Our findings provide strong evidence for the Moral Partiality Hypothesis: in each study, participants said they should protect close others more than distant others. However, the effect of relationship was consistently weaker for "should" judgments than "would" judgments, revealing that people show *relatively less* partiality in their judgments of what is morally right, compared to judgments of how they would act. These findings suggest that when it comes to difficult moral decisions about close others, people think that they would fail to do what is by their own standards "right."

These studies provide further evidence that judgments about moral rules versus decisions about how to act may be sensitive to different considerations (Pletti et al., 2017; Tassy, Deruelle, et

al., 2013; Tassy et al., 2012; Yu et al., 2019). This fits with the growing appreciation of moral cognition as a multifaceted area that involves many distinct kinds of judgments, including whether actions are right or wrong (e.g., Cipolletti et al., 2016; Greene et al., 2009), what to do in hypothetical, lab-based situations (e.g., Bostyn et al., 2018; FeldmanHall et al., 2012; Francis et al., 2016; Patil et al., 2014), what to do when faced with real-world moral choices (Hofmann et al., 2014; Johnson & Goldstein, 2003; Rosenbaum, 2009), and how to evaluate people's moral character (Siegel et al., 2017, 2018).

Our findings also further inform discussions about how relationships and other elements of identity and context influence moral judgments (Hester & Gray, 2020; Schein, 2020). The present work provides additional evidence that relationships influence moral decision-making, and that people think close relationships carry moral weight and generate moral obligations. However, it also shows that different kinds of moral judgments are influenced by relational considerations to different degrees. Further exploring how relational and contextual factors influence a diverse set of moral judgments will help us gain a fuller understanding of the real-world functioning of moral cognition. On a practical level, our findings reinforce an important methodological lesson for moral psychology researchers: it matters both conceptually and empirically what kind of judgment we ask participants to make, and we cannot assume that there will be no difference between what people think they would and should do (see also Barbosa & Jimenez Leal, 2017).

Our findings also are relevant to a number of topics of interest to researchers working at the intersection of psychology and philosophy. Researchers studying hypocrisy—particularly as it relates to inconsistency, self-deception, and akratic thinking—may be particularly interested in our finding that people willingly admit they do not think they would do what they think they should (Alicke et al., 2013; Bartel, 2019; Batson et al., 1999; Jordan et al., 2017; Laurent & Clark, 2019; Lönnqvist et al., 2014). Our findings may shed light on how laypeople think about the relation between moral and practical reasons for action—a topic that ties into many debates in philosophical ethics and action theory. The present work is also relevant for researchers engaged in the philosophical moral partiality debate, and holds particular promise for philosophers who take a more naturalistic and empirically-informed approach to their theorizing (even if the empirical documentation of folk intuitions about moral dilemmas cannot on its own settle normative philosophical questions).

6.1. Future Directions

An important question that we did not address concerns the mechanisms underlying the differences between "would" and "should" judgments, and between responses for close and distant

others (especially for "should" judgments; see Weidman et al., 2020, for mechanisms underlying differences in what people would do).

First, future work should explore what leads people to say they *morally should* protect close others more. One possibility is that when close others are involved, one has to consider competing virtues of *justice* and *loyalty*. Thus, people's normative endorsement of moral partiality could reflect the belief that loyalty is an important moral virtue, and that close relationships give rise to special moral obligations. People might also feel more empathy toward close others than distant others (as they do for ingroup members; see e.g., Gutsell & Inzlicht, 2012; Tarrant et al., 2009). This could lead people to incorporate an expanded set of moral considerations beyond just justice (Batson et al., 1995), could make them more inclined to believe in the underlying goodness of close others, and/or could lead them to assume that the person had a good reason for acting as they did—all of which might lead people to judge the need for (legal) punishment to be less pressing. Finally, attentional mechanisms may contribute to this effect: Berg et al. (2021) show that people tend to focus on the *actor* when a close other transgresses, but the *crime* when a distant other does so. These differences in focus could affect not only the decisions people make (Berg et al.'s focus) but also their conclusions about what is the right thing to do.

These empathic processes, character attributions, and attentional mechanisms might also be mechanisms that underlie the would/should discrepancy. For example, people's empathy toward a close other might make it more challenging for them to do what they think justice requires in these situations (Batson et al., 1995). Another potential explanation for the would/should discrepancy involves reputational concerns and the relational stakes of acting impartially toward close others. People sometimes say that making the morally right decision makes one a worse friend or spouse (Everett et al., 2018), and that leaders who do the morally right thing are judged as cold and lacking empathy (Uhlmann et al., 2013). Thus, people who say they would not do what they should, might be worried about reputational or relational costs of doing what they believe to be right. However, this does not seem to explain why people think they would protect more than they *all things considered* should protect (Study 3)—unless they think that these practical reasons would influence their actions in a manner that is not justified. Our findings also suggest that people think they would protect *distant* others more than they should (though the effect was smaller than for close others). This could be driven by fear of retaliation or worries about the consequences of engaging with law enforcement. These numerous possibilities emphasize that there is much future research to be done to disentangle all of these possible mechanisms and their implications.

Another important future question is whether people *actually* act in real-life scenarios how they *think* they will in hypothetical ones. On the one hand, some work suggests that people may actually behave more morally than they think they will (Teper et al., 2011), positing that affective forecasting failures can lead to poor behavioral predictions. Thus, perhaps people are actually more likely to report a close other (i.e., act as they say they should) than they anticipate. On the other hand, it might be that people are actually rather successful affective forecasters in this domain, if it turns out that thinking about a close other's transgression is highly emotional. Though the general emotionality of these situations has not been directly tested, Weidman et al. (2020) show that psychological distancing—a common strategy for regulating emotions—makes people less likely to say they would report a close other. Given these competing predictions, the relation of "would" and "should" judgments to actual behavior marks an exciting area for future research.

6.2. Concluding Comment

There are several key takeaways from the present findings. First, we have demonstrated that people believe it is *morally right* to treat close others differently than distant others. This suggests that relationship is a powerful factor in folk ethical theory. Second, we have shown that—despite their preference for moral partiality—people also believe that they *would* protect close others more than they *should*, suggesting that people believe they are likely to fail to do what they think is right in moral decisions involving close others. Third, our work has relevance to a number of topics of interest to philosophers, including the relation between moral and practical reasons in folk psychology, and hypocrisy, inconsistency, and akratic thinking. Our results carry both theoretical implications for understanding how relationships influence moral judgments and how different kinds of moral judgments relate to each other, and practical methodological implications for moral psychologists investigating different kinds of moral judgments. Overall, these findings reinforce the claim that decisions involving close others remains a lively domain and fruitful area for moral psychological research.

Chapter 4

How Relationship Affects Adolescents' Decisions to Report Moral Transgressions[75]

1. Introduction

Imagine an eighth grader passing an empty classroom during lunch one day. As she walks by, she sees another student stealing money from someone's backpack. Later, a teacher approaches her and asks whether she saw anything related to the theft. Would she report the student to the teacher, given the opportunity? What if the thief was her very best friend?

Work speaking to this puzzle comes from two different research domains, which have historically seen little contact with one another. In developmental psychology, there is a rich literature exploring how children (Kenward & Östh, 2012; McAuliffe et al., 2015; Riedl et al., 2015; Robbins & Rochat, 2011; Vaish et al., 2011; Yudkin et al., 2020) and adolescents (Brank et al., 2007; Chiu Loke et al., 2011; Friman et al., 2004; Syvertsen et al., 2009; Watson & Valtin, 1997) respond to others' moral transgressions. Alongside this literature is a growing body of work demonstrating that relationships dramatically affect adults' moral decisions (Everett et al., 2018; Hughes, 2017; McManus et al., 2020), including how they respond to others' transgressions (Berg et al., 2021; Lee & Holyoak, 2020; Soter et al., 2021; Waytz et al., 2013; Weidman et al., 2020). Yet little prior work has systematically examined the intersection of these literatures –how adolescents' relationship to a transgressor affects their responses.

The absence of work speaking to this question is striking, as adolescents who witness a close friend acting immorally face a particular dilemma that does not arise when the transgressor is a stranger or distant acquaintance. On the one hand, children and adults alike are strongly motivated to punish wrongdoing (Graham et al., 2009; Hofmann et al., 2018; Schein & Gray, 2018). On the other hand, people have a fundamental tendency to protect the wellbeing of close others (Aron et al., 1991; Hamilton, 1964). This tension may be heightened in adolescence, which is characterized as a time when peer relationships are an particularly central part of one's social world (Berndt, 1982; Brown & Larson, 2009; Hart & Carlo, 2005; Hunter & Youniss, 1982), and moral development undergoes

[75] The in preparation version of this paper is co-authored with Martha Berg, Ethan Kross, and Susan Gelman. Thanks also to Annalee Miklosek for her help in coding open-ended responses.

substantial growth (Gibbs, 2019). Thus, when the immoral actor is a close other, an adolescent must decide whether to prioritize justice or loyalty. In the present work, we seek to fill this gap in the literature, examining how relationship to a transgressor affects adolescents' decisions to report an immoral act.

1.1 Reporting Transgressions in Childhood

From at least three years old, children show strong motivation to respond to others' immoral acts (Kenward & Östh, 2012; McAuliffe et al., 2015; Riedl et al., 2015; Robbins & Rochat, 2011; Vaish et al., 2011; Yudkin et al., 2020), a tendency that continues into adulthood (Graham et al., 2009; Hofmann et al., 2018; Schein & Gray, 2018). Although there are many possible ways to respond to another's transgression, including intervening to prevent the act (e.g., Vaish et al., 2011), helping the victim (e.g., Lee & Warneken, 2020; Riedl et al., 2015), or punishing a transgressor by taking away desired resources (e.g., Kenward & Östh, 2012; McAuliffe et al., 2015; Yudkin et al., 2020), one important strategy is *reporting* the wrongdoing to an authority figure. This is an especially relevant response for children and adolescents, who may not be positioned to directly punish others or rectify harm done, but who frequently have access to an authority figure such as a parent, teacher, or police officer. Thus, reporting behavior, sometimes referred to as "tattling," has received much attention in the developmental literature (Chiu Loke et al., 2011, 2014; Den Bak & Ross, 1996; Heyman et al., 2016; Ingram & Bering, 2010; Lyon et al., 2010; Misch et al., 2018; Ross & Bak-Lammers, 1998; Watson & Valtin, 1997).

The decision to report a transgressor can carry significant social weight, and may explain age differences in children's willingness to report others. Whereas children from three to seven years old are quite willing to report both peers and adults who commit major and minor transgressions (Chiu Loke et al., 2011; Den Bak & Ross, 1996; Heyman et al., 2016; Ross & Bak-Lammers, 1998), by age eleven children endorse reporting only major transgressions (Chiu Loke et al., 2011, 2014). A primary hypothesis for this reluctance to report as children enter adolescence is the increase in the perceived social costs of reporting (Friman et al., 2004). Adolescents are more concerned than younger children about their peers' perceptions (Jankowski et al., 2014; Juvonen & Murdock, 1995; LaFontana & Cillessen, 2010; Mulvey & Killen, 2016; Sebastian et al., 2008), and about negative consequences in general (Vasey et al., 1994). Consistent with this hypothesis, adolescents' willingness to report potential violence is mediated by their beliefs about unfavorable consequences of confiding in a teacher (Syvertsen et al., 2009), and increases under conditions of anonymity (Brank et al., 2007). Further,

these reputational concerns appear to be grounded in reality: Friman et al. (2004) documented strong negative associations between perceived tattling and both social status and likeability for adolescents.

1.2. Adult Reporting and Relationships

In the domain of adult research, a growing body of work has demonstrated that social relationships dramatically affect how adults respond to witnessing someone commit a moral transgression (Berg et al., 2021; Lee & Holyoak, 2020; Waytz et al., 2013; Weidman et al., 2020). After witnessing someone transgress, adults are far more likely to report the transgressor to an authority figure when the transgressor is someone they barely know (e.g., mail carrier, dentist) than when it is one of the people closest to them (e.g., best friend, sibling)—especially when the transgression is a severe one (Berg et al., 2021; Weidman et al., 2020; see also Waytz et al., 2013). Further, adults say that it is *morally right* to report close others less than distant others (Soter et al., 2021).

1.3. Relationships in Adolescence

Despite this strong evidence that relationships impact reporting decisions in adults, less is known about the developmental trajectory of this pattern. Though prior work has explored how children's judgments about transgressions vary based on their relationship to the *victim* of a transgression (e.g., Costin & Jones, 1992; Olson & Spelke, 2008; Recchia et al., 2013; Slomkowski & Killen, 1992; Smetana & Ball, 2018), less work has directly explored children's and adolescents' relationship to the *transgressor*. Specifically, little is known about how this influences their reaction to the transgression—especially whether they decide to report to an authority figure. In the existing work on children's (ages 3-12) reporting decisions, there has been considerable variation across studies in whether the transgressor was a stranger or anonymous (Chiu Loke et al., 2011, 2014; Heyman et al., 2016; Vaish et al., 2011), a hypothetical or real friend or sibling (Den Bak & Ross, 1996; Ross & Bak-Lammers, 1998; Watson & Valtin, 1997), or a classmate (with relational closeness either unspecified or not controlled for, Friman et al., 2004; Ingram & Bering, 2010; and most of the questions in Brank et al., 2007). Relationship to the transgressor has thus varied across studies, but not been systematically manipulated. The few studies that have directly manipulated relationship via manipulation have done so only in younger children (e.g., ages 4-9) and have examined ingroup/outgroup relationships (Misch et al., 2018) or relationships to adults (Lyon et al., 2010), rather than individual peer relationships. Moreover, such studies have typically examined how relationship affects children's assessments of the transgressor, but not how it influences their responses and reporting decisions (Peets et al., 2007).

Yet as briefly noted earlier, adolescence in particular is a time when peer relationships are central to children's social lives (Berndt, 1982; Hart & Carlo, 2005). In this period, friends begin to

exert a significant influence over one another (Brown & Larson, 2009; Youniss & Haynie, 1992) and become a crucial element of social adjustment and belonging (Bukowski et al., 1993; Rubin et al., 2004; Waldrip et al., 2008). Given the importance of these peer social ties, relationship closeness may be a salient factor in determining how adolescents make social decisions. This is especially plausible given the evidence discussed above emphasizing that social consequences appear to be a major consideration influencing whether adolescents choose to report (Brank et al., 2007; Friman et al., 2004). Reporting one's best friend surely carries much heavier potential social costs than reporting a distant acquaintance: it risks, for instance, damaging a highly valued relationship and being perceived as a bad or disloyal friend. Indeed, in the domain of bullying, it has been found that friendship with a bully is often a motivational factor that discourages bystander intervention (Forsberg et al., 2014; Thornberg et al., 2012, 2018).

To our knowledge, only one study to date has systematically examined how relationship affects adolescents' decisions to report transgressions to an authority figure. Brank et al. (2007) investigated middle schoolers' responses to a classmate bringing a weapon to school and found that students indicated slightly more reluctance to report a friend than an unspecified classmate. However, Brank et al.'s (2007) evidence came from a single item embedded in an extensive questionnaire exploring the highly specific (and emotionally charged) issue of students' responses to weapons at school. Thus, more systematic research is needed.

2. The Present Study

The goal of this research is to test whether and how relationships influence how adolescents respond to witnessing a moral transgression. To address this issue, we asked a large sample of sixth through ninth graders to imagine witnessing either their best friend, or a student they know distantly, committing a high or low severity theft. We then asked whether they would report the act to an inquiring teacher. Our participants were recruited from urban schools with predominantly Black and low-income populations, allowing us to conduct this research with participants who are often underrepresented in empirical psychological research. We focused on theft-based transgressions for three reasons: thefts are transgressions that adolescents realistically might do or observe; theft easily allows for manipulation of severity, based on whether a high- or low-value item is stolen; and focusing on theft allows for straightforward comparison with existing adult work in this domain (e.g., by Weidman et al., 2020).

Based on adult studies documenting a reluctance to report close others who transgress, and adolescent research showing the heightened importance of peer relationships at this point in

development, we hypothesized that adolescents would be more likely to report a distant classmate who transgresses, than their best friend, and more likely to report severe transgressions than minor ones. We preregistered the prediction that this effect would be more pronounced for high severity transgressions, based on the available evidence from adult studies; however, we also noted that we did not have a strong prediction regarding the relationship by severity interaction, as this was the first study investigating this phenomenon in adolescents.

This study was preregistered through the Open Science Framework; blinded preregistration for peer review available in "Files" at https://osf.io/znstf/?view_only=02c0894e02aa416aad618e8493310a4e. All methods and analyses follow the preregistered plan, except where otherwise noted.

3. Method

3.1. Data Collection

This investigation was part of a larger data collection effort that included a variety of studies designed by scientists affiliated with Character Lab Research Network (CLRN). CLRN simultaneously rolled out multiple independent studies, and students were randomized to one of the studies running in their school. This study was conducted on school computers during class time in participating schools over the course of a two- to three-week testing window. On a predetermined testing day, a teacher proctor at each school administered the CLRN research activities to students. To introduce the study, teachers read a script that explained to students that all research activities were part of an educational research initiative at their school, that participation was voluntary and they were not being graded, and that teachers would not see their answers. Teachers also instructed students to focus on their own computers and (if relevant) not to look at classmates' screens. Upon logging into the CLRN platform, all students first viewed an assent screen that reiterated this information and, in addition, explained that parents would not see their responses and that their names and any other unique identifying information would not be shared with researchers. Students who agreed to participate were then directed to the survey.

3.2. Participants

Our participants came from CLRN Stratum D: mostly Black, high number of students receiving free/reduced-price lunch, often urban schools, in the United States. These schools represent 12% of all middle and high schools in the US. The full sample consisted of 1,349 sixth through ninth grade students. Participants were excluded from analysis for the following criteria: not finishing the survey (n = 213), skipping over more than two (out of eight) dilemmas (n = 4), selecting "no" when

asked if they answered honestly (n = 44), saying they were distracted during the survey (n = 149), and for self-reporting ages younger than 10 or older than 16 or not reporting an age (n = 8; not pre-registered), for a final adolescent sample of N = 913.

The final sample consisted of 39% sixth graders, 27% seventh graders, 26% eighth graders, and 8% ninth graders; mean age was 12.36 (SD = 1.14; range 10-16). Fifty-five percent identified as girls, 44% as boys, and 0.12% as other. The sample was 2% Asian, 7% bi-/multi-racial, 63% Black/African American, 13% Hispanic/Latino, 1% Native American, 0.1% Native Hawaiian/Pacific Islander, 7% white, and 7% other.

3.3. Procedure

Our paradigm was adapted for a school context and younger sample from Weidman et al. (2020). We followed Weidman et al. (2020) in focusing on theft-based transgressions of high and low severity. In an online Qualtrics survey, participants were first asked to imagine their best friend at school, and a student in another class whom they do not know very well. For each relationship, they were asked to think of a specific person and hold that person in mind for the rest of the survey. Next, participants were presented with eight vignettes. In each, they were asked to imagine that they witnessed either their best friend, or the distant student, committing a theft. Half of the thefts were high severity (e.g., taking the teacher's laptop) and half were low severity (e.g., taking an extra cookie from the cafeteria without paying). Severity was distinguished by the value of the stolen item; we verified the perceived severity of each theft in a pilot study with adult participants. The full list of scenarios is included in Appendix F. Severity and relationship were fully crossed: each participant responded to two dilemmas with each relationship/severity combination. Participants were then asked to imagine a teacher approaching them and asking whether they had seen anything suspicious. (Weidman et al. (2020)'s original studies involved reporting to a police officer.) Participants then indicated whether they would report the transgressor to the teacher, on a 6-point Likert scale (1 = "Definitely would not report"; 6 = "Definitely would report"). After the final trial only, participants were asked to write what they were thinking about as they responded to the preceding question (the report measure for the final vignette). Open-ended data are currently being coded, and results from these analyses are not reported in this dissertation. Finally, participants reported their level of trust in teachers (0 = "Very untrustworthy"; 6 = "Very trustworthy") as an exploratory measure, and completed a set of demographic questions.

4. Results

Overview of Analyses

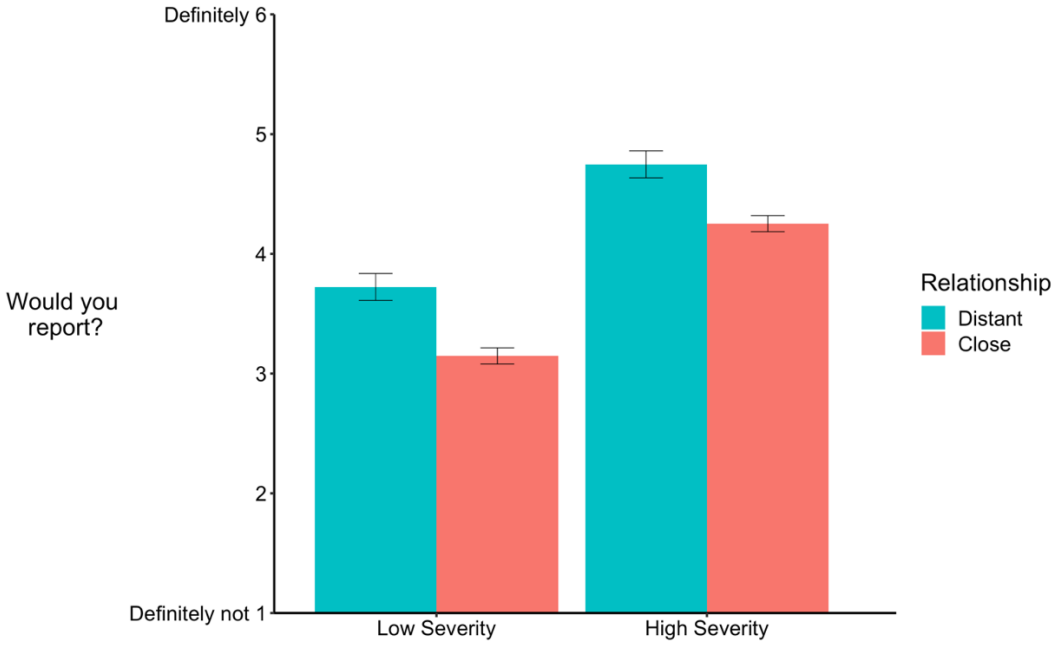
To test the effects of relationship and severity, we ran a mixed linear model using the lme4 package in R (Bates et al., 2015). We included participant and dilemma as random intercepts; thus, variance across both individual participants and differences in vignettes were accounted for in the model. For random slopes, severity and relationship were allowed to vary by participant, and relationship was allowed to vary by dilemma (the maximal model that reached convergence). A likelihood-ratio test indicated that this model including random slopes provided a better fit for the data than a model that only included random intercepts ($\chi^2(7) = 324.85, p < .001$). Binary fixed effects (relationship and severity) were contrast coded (-0.5/0.5).

Main Results

Participants were less likely to report close others ($M = 3.70$; $SD = 1.76$) than distant others ($M = 4.24$, $SD = 1.83$; $b = -0.54$, $t(6.76) = -8.4$-, 95% CI = [-0.68, -0.40], $p < .001$). Participants were also less likely to report low severity ($M = 3.44$; $SD = 1.78$) than high severity transgressions ($M = 4.50$; $SD = 1.69$; $b = 1.06$, $t(6.73) = 10.33$, 95% CI =[0.82, 1.30], $p < .001$). There was no significant relationship by severity interaction ($b = 0.08$, $t(4.96) = 0.68$, 95% CI =[-0.23, 0.39], $p = .53$). Results are displayed in Figure 4.1.

Figure 4.1

Reporting Close versus Distant Others for High and Low Severity Transgressions



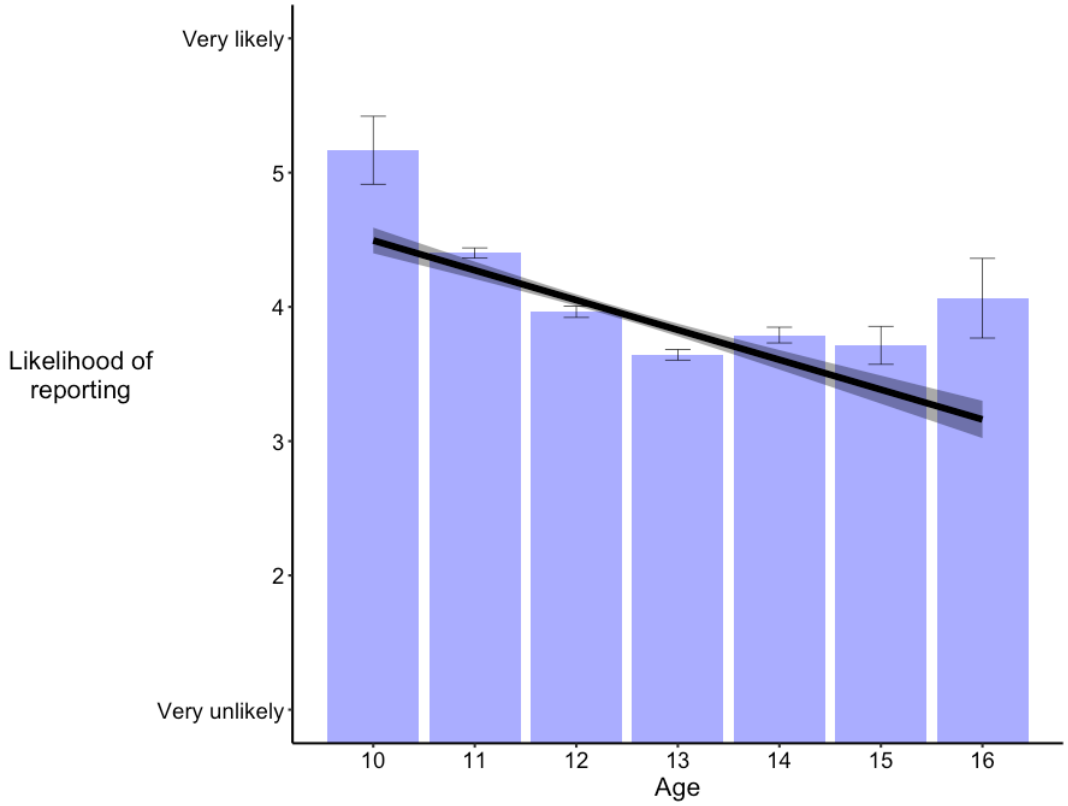
Note. Error bars show standard error of the mean.

Exploratory Analyses

We conducted three additional exploratory analyses to check for age and gender effects, and to control for baseline trust in teachers. Reporting decreased overall with age (b = -0.22, t(910.82) = -6.20, 95% CI = [-0.30, -0.14], p < .001). Main effect of age is displayed in Figure 4.2; age did not significantly interact with relationship or severity. Girls were more likely to report overall than boys (excluding non-binary participants, n = 11, for this analysis; b = 0.53, t(899.51) = 6.45, 95% CI = [0.37, 0.69], p < .001; random slopes of relationship and severity over participant removed to avoid convergence failure). Finally, higher teacher trust (M = 3.32, SD = 1.17) predicted higher reporting (b = 0.42, t(902.6) = 12.90, 95% CI = 0.36, 0.48], p < .001; random slope of relationship over participant removed to avoid convergence failure). None of these additional parameters eliminated the main effects of severity and relationship.

Figure 4.2

Reporting Decreases with Age



Note. Error bars show standard error of the means.

5. Discussion

Our driving question for this study was: (how) does relationship to a transgressor affect whether adolescents decide to report a peer whom they witness stealing something? This question emerges from the intersection of two lines of research: the developmental literature on adolescent reporting decisions, and work on how relationship affects adults' responses to moral transgressions. We studied this question in a predominantly Black and low-income adolescent population, a demographic that is underrepresented in research on the development of moral reasoning.

We found that adolescents were indeed more reluctant to report a best friend who stole something than a distant classmate. We also found that adolescents were more likely to report both close and distant others who had committed more serious transgressions, than those who had committed minor ones—again consistent with findings in adults. Additionally, we found that older participants in our sample were overall more reluctant to report than younger participants. This age effect is consistent with previous developmental work showing decreasing reporting with age in 6-11 year-olds (Chiu Loke et al., 2011); our work documents the continuation of this trend into early adolescence. This finding is also consistent with work suggesting that as young people enter late childhood and early adolescence, they gain the capacity for more complex moral reasoning—moving beyond the simple application of moral rules, and towards considering multiple perspectives, thinking more abstractly, and integrating more complex information processing strategies (e.g., see Graber et al., 2018; Simmons & Blyth, 2017 for discussion).

Notably, we did not observe a severity by relationship interaction. Although participants were more likely to protect close others overall, the degree to which they showed this preference did not depend on the severity of the theft. This marks a striking difference from the pattern seen in adults (Weidman et al., 2020), whose preferential protection of close others is consistently greater for severe transgressions. This could indicate a developmental difference: the fact that we see an equally strong reluctance to report close others, even for low-severity thefts could suggest that relationships have a more robust effect for this age group, and that discretion in when to prioritize not reporting close others develops later in adolescence or early adulthood. (See Supplemental Study in G for a study that provides preliminary data addressing this question.) This lack of discretion in when to avoid reporting close others—i.e., not doing so even when consequences are likely to be relatively low—could also reflect the central role that friendships occupy in adolescents' social worlds.

However, there are several differences between the present study and existing adult work, such that we cannot be certain that this lack of interaction reflects developmental differences. The present

adolescent sample differed demographically from the adult samples tested in Weidman et al. (2020)'s work: our sample was predominantly Black and low-income, whereas previous samples have been majority white and relatively higher socioeconomic status. These racial and socioeconomic differences might reflect relevant differences in how participant groups view and respond to the relevant authority figures, as well as the consequences of their decisions (e.g., Gregory & Ripski, 2008; Panditharatne et al., 2021; Tuma & Livson, 1960; Viki et al., 2006; Wright & Weekes, 2003). In the present study, participants imagined reporting the transgression to a teacher, rather than a law-enforcement officer; given that teachers might be assumed to know all the students involved, participants could have been motivated not to report a theft even when the consequences would otherwise likely be mild. Future work focused on the emergence of the difference in reporting close others for high or low severity transgressions should control for these factors.

Another key question for future research concerns at what age children begin to show this reluctance to report close others who transgress—or whether this tendency is present throughout development. Because adolescence is a particularly transformational time for the role of peer relationships and the importance of friendships, perhaps adolescence marks the beginning of this pattern of preferential non-reporting (but see Misch et al., 2018).

The mechanisms underlying the reluctance to report close others across age groups is also an important area of future study. In the adult literature, feelings of loyalty to close others and concerns about harm to close others have both been proposed as important mechanisms explaining the bias towards protecting close others, as have self-interest and reputational concerns (Weidman et al., 2020). Adolescents' reluctance to report may similarly be motivated by both self-interested reputational concerns about social standing, or by relationship dynamics such as loyalty and desire to protect close friends. There may also be more cognitive-developmental mechanisms in play, such as increased capacity for perspective-taking (especially of close others), and ability to apply moral rules in a more nuanced and selective way. Exploring the developmental trajectory of these concerns and capacities has the potential to provide a richer understanding of the motivations behind adolescents' responses to the transgressions of those around them (and will offer key insight into how the reluctance to report can be reduced, when appropriate).

Finally, future work should explore the generalizability of these findings to other kinds of transgressions. Though this pattern has been documented across multiple domains in adults (both theft and sexual assault transgressions, Weidman et al., 2020), the data presented here cannot directly speak to how adolescents may respond to transgressions in other domains. However, work on

adolescent responses to bullying do suggest that relationship to an aggressor can be a powerful force preventing bystander intervention (Forsberg et al., 2014; Thornberg et al., 2012, 2018), offering some initial evidence that a reluctance to intervene when a transgressor is a close other extends beyond the domain of theft.

Although there remains much to investigate in this domain, the present findings establish that interpersonal relationships are an important factor in understanding how adolescents respond to the moral transgressions of others. In the rich literature on children's and adolescents' reporting decisions, almost none has systematically explored this dimension previously. We thus hope that the present study will spur future research into how different kinds of relationships—and other relevant social factors—influence how people at various stages of development judge and react to others' moral transgressions.

Concluding Remarks

I want to conclude this dissertation by offering some thoughts on future work. In both my philosophical and empirical projects, there remain many open questions, and I take the chapters in this dissertation to be starting points rather than end points. In what follows, I outline some of the lingering questions and plans for future research that have stemmed from the present projects.

1. Future Plans for the Philosophical Project

My work on acceptance lays the foundation for future research into the nature and ethics of doxastic regulation. My planned projects in this domain include both philosophical and empirical work.

My immediate next planned project takes the broad methodological approach I use in my work on acceptance—that we can learn a lot about doxastic control by looking to empirical research on other kinds of mental state control—and uses it to engage more directly in the debate over doxastic (in)voluntarism.[76] The existing voluntarism debate has focused nearly exclusively on whether we can control belief *formation*. Call this "front-end" control. The general consensus is that we cannot, or at least not directly, and at least not in responsive to non-evidential reasons.[77] However, changing and updating in response to evidence is just one of the primary functional roles of belief. Beliefs also serve as the default guiding basis for a wide variety of cognitive, reasoning, and action processes; they are powerful prepotent mental states that cause effects across a wide range of psychological mechanisms. In virtue of this, I plan to argue that we can exert significant "back-end" control over this guiding or steering role of beliefs—and that we know a lot about what this kind of control looks like from research on cognitive control mechanisms in controlling mental states like emotions, cravings, and other thoughts. Drawing on work from authors such as Sripada (2021) and Bermúdez (2021), I will ultimately argue that the control profile at hand is one of *self-control*. An upshot of this view is that evidentialist theories of belief, which care primarily about belief's evidence-responsive capacities, tend to be more involuntaristic, while pragmatist or dispositionalist theories of belief, which care more

[76] This debate concerns whether and to what extent we can control what we believe or choose our beliefs.

[77] Though authors such as Hieronymi have argued that this fact does not pose a problem for *epistemic* deontology, because we have the capacity to respond doxastically to the right kinds of reasons for belief (i.e., evidence), and in doing so we can manifest our epistemic agency.

about the way belief guides cognition, reasoning, and action, will ultimately be more voluntaristic. This project will also allow me to address a lingering challenge for my present account of acceptance: as I currently discuss acceptance, I insist that the "underlying belief" remains untouched by our response modulation efforts. However, on some prominent theories of mental states, if you completely cut off a mental state's functional role, it's hard to say that you still *have* that mental state. This two-pronged approach to belief and doxastic control gives us the resources to handle this worry: we can say that the front-end appraisal of the evidence remains unchanged by the back-end control mechanisms—but then whether to call the mere appraisal a *belief* may depend in part on what kind of theory of belief one goes in for.[78]

The preceding paper will be primarily a descriptive project. In another paper, I will explore how we ought to think about the normative landscape of these two functional roles of belief. Much recent work has focused on theorizing about the relationship between epistemic norms and moral or practical norms. For instance, advocates of encroachment theories propose that the moral content or stakes of a belief can affect whether an agent is epistemically justified in holding that belief. I will use my two-pronged framework (i.e., the explicit distinction between the front-end and back-end functional roles of belief) to offer a new way of framing the possible ways in which these different kinds of norms can interact with each other to govern in belief. I will argue that we should think of familiar evidentialist/purist epistemic norms as really being "front-end" doxastic norms, but that what such norms say about "back-end" belief guidance is far less clear—and that in this back-end role, there is much more room for practical and moral considerations to get purchase. My project will in part be to lay out the choice points available for theorists who want to understand how various kinds of norms interact with belief, but I will also argue more positively that appealing to moral and normative constraints on back-end belief guidance is the most promising route to getting what advocates of encroachment or doxastic wronging want from their views, but that carving up the space in this way avoids many of the theoretical challenges that plague those kinds of views.[79]

I also intend to use my accounts of acceptance and doxastic regulation to make progress on other specific debates within the ethics of belief and philosophy of mind. In some instances, I will continue to argue that regulating beliefs in the way I prescribe is a good thing—such as cases involving

[78] In fact I think it is a benefit of this framework that I don't have to commit myself to any specific account of what a belief actually is: by laying out the functional roles and mechanisms in play, I can say what I want about control, and then people who have other reasons to prefer one kind of theory over another can do with the terminology as they like.
[79] I also think that some doxastic wronging advocates have inflated the front-end and back-end functional roles of belief in a way that this framework will help significantly clarify.

involving profiling, stereotyping, or generic reasoning about social groups. In these cases, an agent has (by stipulation) strong statistical or group-level evidence about people of a particular kind of social group—but she also seems to have strong moral reason not to let that assessment of the evidence play the role it usually would in guiding her reasoning, cognition, and action. Though some work has already been done to apply acceptance to cases of racial generalizations (e.g., Bolinger, 2020), I think there is more to be said by leveraging my psychological account, including addressing: the ways in which these beliefs naturally license and lead to—but do not necessitate—certain patterns of thought and inferences; why prior work has struggled with whether to characterize these kinds of considerations and acceptances as epistemic or merely moral; and the cognitive difficulty of exercising this kind of control, which I think is important for our moral assessments of agents' mistakes. There are also plenty of cases in which accepting seems to be important in our distinctively epistemic projects: for instance, much focus has been paid recently to withholding or suspending judgment in inquiry. On my view, a plausible way to cash those ideas out is via the response modulation mechanisms I describe: preventing the appraisal of the evidence—even if we have high confidence—from playing its usual downstream role, in favor of a withholding attitude.

Much of my focus, so far, has been on cases where acceptance is a good thing. But insofar as I think it is a psychological capacity we actually have, it is a skill that can be used poorly as well. I think that my account can help us understand doxastic phenomena such as self-deception and denial. I also think they may play a role in social phenomena such as ideology—especially cases where people seem to learn to habitually and skillfully prevent themselves from responding to even very strong evidence in the usual way. This also raises conceptual questions about the relationship between acceptance and rationality: when should we think an agent is accepting, and when should we say she is actually believing or appraising the evidence irrationally? Does it depend on the kinds of reasons that she actually is, or takes herself to be, responsive to? And (how) does the distinction between front-end and back-end functional roles, control, and norms of belief affect how we even conceptualize questions about rationality? Finally, I think that there is much to say about how we can do real harm to other people when certain kinds of social norms place demands on them to accept things despite good evidence. For instance, part of what we do when we gaslight people, or we tell them not to make an issue out of insulting comments or microaggressions, is that we ask them to do the work of accepting: of not letting their belief (for instance, about the insulter's motivations or values) play its usual guiding role in how they assess and interact with that person.

Having rambled about my philosophical ideas for a while, let me also add that I hope to do various kinds of empirical work that build on my philosophical work on doxastic control and acceptance. I think there are ways I can directly test the account of acceptance I put forward. For instance, we can design experiments that induce beliefs in people (or verify preexisting beliefs), give them tasks that involve making judgments or decisions involving those beliefs, and then instruct some people to accept something else—that is, to prevent those beliefs from guiding their judgments and decisions in the usual way. We can then look for classic markers of inhibitory control mechanisms, such as slower reaction times and increased failure under cognitive load.

More broadly, I want to develop research exploring how laypeople think about control over mental states, and how those conceptions of control affect their moral judgments their own and others' mental states. With Ethan Kross and Susan Gelman, I am already beginning a line of work exploring how people think about two possible kinds of control over mental states like thoughts and emotions: a) controlling whether an unwanted mental state *arises in the first place*, and b) controlling whether and how you continue to elaborate on that mental state. Prior work on folk theories of mental state control has not been sensitive to this distinction. Our data so far suggest that nearly half of our participants thought that people should be able to control *whether an unwanted thought or emotion state even arises*. We are currently exploring how individual differences in this "spontaneous control" affects people's moral judgments about unwanted mental states in others, and how it affects their responses to their own unwanted mental states. I am excited to continue working on these projects.

I think there is also much work to do to understand the folk concept and ethics of *belief*, specifically. For instance, what little work exists in this domain suggests that the folk concept of belief is surprisingly (from a philosophical perspective, anyways) voluntaristic (e.g. Cusimano & Goodwin, 2019; Cusimano & Lombrozo, 2020). I want to further probe what this voluntarism actually amounts to. In particular, my suspicion is that when laypeople think about belief, they don't think of the narrow, front-end evidentialist conception that many philosophers have in mind—but rather, they also think of belief as an attitude that richly guides our cognition, reasoning, and action. If that's right, then perhaps we can push people's judgments around: perhaps when people think more about belief as a state that is evidence-responsive (i.e., front-end functionality) they will become less voluntaristic, and when they think about belief as a guiding or steering mental state (i.e., back-end functionality) they will become more so. I think this is especially plausible because often what we actually *care* about when we think about others' beliefs is the way those beliefs guide their mental and behavioral lives—which is why we do so often evaluate people robustly in terms of what they believe.

2. Future Plans for the Psychological Project

There is also much future work to do stemming from the empirical chapters of this dissertation. (Though I have already highlighted some of my current psychological work in the preceding section.)

In one future project, I hope to build on the findings from the final study in Chapter 3, that showed people's ability to distinguish between what they *morally* ought to do, and what they *all-things-considered* ought to do. Philosophers differ dramatically in how they think moral reasons do and should interact with other kinds of reasons for action. To my knowledge, little work has explored how laypeople think about this. Do people—like some philosophers—think that moral considerations are, or should be, overriding? Or do they think—like other philosophers—that moral considerations enter our deliberations on exactly the same footing as any other kind of practical consideration? Or do they occupy some middle role? A related sries of questions might investigate what kinds of things people take to fall into the domain of the moral in the first place. (Relationships are an interesting case: do people think that special considerations regarding close others are *moral* considerations, specifically, or some other kind of practical consideration?)

Another line of ongoing research, led by Yeonjee Bae, is exploring how partiality in responding to transgressions develops in children as young as four years old. The developmental trajectory of partiality versus impartiality may have interesting philosophical implications. For instance, do we learn how to be impartially moral by starting out partial and learning to expand our moral circle? Or do we start by valuing everyone equally, and learn about special relational obligations as we get older?

I am excited to continue to explore how various other kinds of relationships influence moral cognition as well. I'm interested in the role of reciprocity expectations in relational moral cognition: how does what people expect close others would do in a situation influence their own choices? Another relatively less explored relationship in empirical moral psychology is our relationship to ourselves. In an ongoing project with Fan Yang, I am investigating whether people think there are moral duties to the self, and how they reason about those duties in comparison to moral duties to others. (The data so far strongly suggest that people *do* think there are moral duties to the self, and that those are distinct from social conventions or duties to others.) In my post doc, I will be working on intergroup relationships and moral cognition. One question I have been mulling over lately, and hope to be able to explore at some point, is whether there are differences in how we reason about groups marked by clear physical differences (e.g., race, or class as marked by clothes) versus groups marked by *beliefs* (e.g., political party). Do we tend to run these two together in our minds—for

instance, do we infer that that people in different racial groups have different beliefs, and create visual markers for people in different political parties? Or do they enter our reasoning and moral judgment in distinct ways?

* * *

This concluding section has really just been a place for me to document the current state of my idea-space as I am finishing up this doctoral program. There is, in summary, plenty more work to do. Of course, the suggestions here are all just possibilities; I'm sure some projects will pan out more successfully than others, and that new projects will develop from and instead of the ones described here. Regardless, I am so grateful to have been able to get the training over the past five years that will allow me to do it, and I'm so excited for all the research ahead.

APPENDIX A

List of Chapter 3 Moral Transgressions

Below is the set of theft dilemmas used in the studies presented in Chapter 3.

1. Stealing a wallet from the seat of a parked car.
2. Stealing an unattended laptop in a coffee shop.
3. Taking a dog tied up outside a shop.
4. Stealing jewelry from a locked store display.
5. Blackmailing someone for money by threatening to post unflattering pictures of them online.
6. Breaking into a house and stealing a TV.
7. Taking money out of a charity donations basket.
8. Taking a credit card left on a restaurant table.

APPENDIX B

Study 3 Descriptive Statistics

Table B.1 displays the cell descriptive statistics for Study 3.

Table B.1

Mean Protecting across Study 3 Judgment Conditions (1-6 Scale)

	Ideally Should	**Morally Should**	**Overall Should**	**Would**
Close	2.49 (1.61)	2.49 (1.74)	3.12 (1.71)	3.91 (1.77)
Distant	1.78 (1.24)	1.75 (1.30)	2.00 (1.35)	2.42 (1.60)
Total	2.14 (1.48)	2.12 (1.58)	2.55 (1.64)	3.17 (1.84)

Note. Standard deviations in parentheses.

APPENDIX C

Chapter 3 Supplement 1: Police Trust Models

In each of our four studies, we collected a self-report measure of police trust from participants. In the main text we note that controlling for police trust did not change the pattern of primary findings in any study. Here we report full statistics for the models including police trust as a covariate. One general finding emerged: across all studies, higher police trust predicted lower protecting. For several studies (1a, 1b, 2), this effect was significantly stronger for decisions involving close others.

Table C.1

Study 1a: Model including Police Trust

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Relationship [a]	1.23	.04	1.15	1.31	<.001
Judgment [b]	-0.41	.14	-0.69	-0.13	.004
Police Trust [c]	-0.26	.04	-0.36	-0.16	<.001
Relationship x Judgment	-0.49	.09	-0.67	-0.31	<.001
Relationship x Police Trust	-0.22	.03	-0.28	-0.16	<.001
Judgment x Police Trust	0.01	.09	-0.17	0.19	.90
Rel. x Judgment x Police Trust	-0.04	.06	-0.16	0.08	.49

Note. N = 299. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] -0.5 = distant, 0.5 = close. [b] -0.5 = would, 0.5 = should. [c] Police trust was mean centered.

Table C.2

Study 1b: Model including Police Trust

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Relationship [a]	1.49	.04	1.41	1.57	<.001
Judgment [b]	-0.91	.14	-1.18	-0.64	.004
Police Trust [c]	-0.37	.04	-0.45	-0.29	<.001
Relationship x Judgment	-0.69	.09	-0.87	-0.51	<.001
Relationship x Police Trust	-0.06	.03	-0.12	-0.001	.03
Judgment x Police Trust	0.06	.09	-0.12	0.24	.50
Rel. x Judgment x Police Trust	0.002	.06	-0.16	0.12	.97

Note. N = 316. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] -0.5 = distant, 0.5 = close. [b] -0.5 = would, 0.5 = should. [c] Police trust was mean centered.

Table C.3

Study 2: Model including Police Trust

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Relationship [a]	0.95	.03	0.98	1.01	<.001
Judgment [b]	-1.13	.03	-1.19	-1.07	<.001
Police Trust [c]	-0.29	.04	-0.37	-0.21	<.001
Relationship x Judgment	-0.83	.06	-0.95	-0.71	<.001
Relationship x Police Trust	-0.09	.02	-0.13	-0.05	<.001
Judgment x Police Trust	0.04	.02	0.0008	0.08	.02
Rel. x Judgment x Police Trust	-0.04	.04	-0.12	0.04	.35

Note. N = 356. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] -0.5 = distant, 0.5 = close. [b] -0.5 = would, 0.5 = should. [c] Police trust was mean centered.

Table C.4

Study 4: Model including Police Trust

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Relationship [a]	1.01	.05	0.91	1.11	<.001
Judgment Contrast 1[b]	0.81	.10	0.61	1.01	<.001
Judgment Contrast 2[c]	0.44	.13	0.19	0.69	<.001
Judgment Contrast 3[d]	-0.05	.14	-0.32	0.22	.69
Police Trust [c]	-0.23	.03	-0.29	-0.17	<.001
Relationship x J. Cont. 1	0.62	.11	0.40	0.84	<.001
Relationship x J. Cont. 2	0.02	.13	0.13	0.63	.005
Relationship x J. Cont. 3	0.02	.15	-0.27	0.31	.87
Relationship x Police Trust	0.0008	.03	-0.06	0.06	.98
J. Cont. 1 x Police Trust	-0.22	.06	-0.34	-0.10	<.001
J. Cont. 2 x Police Trust	-0.002	.07	-0.14	0.14	.97
J. Cont. 3 x Police Trust	-0.07	.08	-0.23	0.09	.35
Rel. x J. Cont. 1 x Police Trust	-0.04	.06	-0.16	0.08	.48
Rel. x J. Cont. 2 x Police Trust	0.14	.07	0.003	0.28	.06
Rel. x J. Cont. 3 x Police Trust	-0.01	.08	-0.17	0.15	.89

Note. N = 609. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] -0.5 = distant, 0.5 = close. [b] (-1/3, -1/3, -1/3, 1) contrast: comparing "would" judgments to all "should" judgments. [c] (-1/2, -1/2, 1, 0) contrast: comparing "overall should" judgments to "morally should" and "ideally should". [d] (-1, 1, 0, 0) contrast: comparing "ideally should" to "morally should" judgments. [e] Police trust was mean centered.

APPENDIX D

Chapter 3 Supplement 2: Study 3 Exploratory Model

In this supplement, we report the results of an additional exploratory model run on the data from Study 3.

In this model, we test for a two-way interaction between relationship and judgment type, among only participants in the "overall should" and "actually would" judgment conditions (N = 305). In this model, participants were more likely to protect close others than distant others (b = 1.29, t(302.88) = 16.40, p < .001), and participants overall said that they would report more than they overall should report (b = 0.60, t(302.99) = 3.98, p < .001). There was a significant interaction between judgment and relationship (b = 0.36, t(303.07) = 2.03, p = .02): the difference between judgments was greater for close others (b = -0.78, t(302.99) = 4.22, p < .001) than distant others (b = -0.42, t(302.99) = 2.73, p < .01).

APPENDIX E

Chapter 3 Supplement 3: Blocked Judgment Studies

In this supplement, we present two additional studies run in the course of this research project.

1. Supplemental Study 1

Supplemental Study 1 (S1) was conducted first and designed to test the relation between "should" and "would" judgments in punish-or-protect dilemmas. Afterward, we concluded that the methodology was flawed, as we lacked a manipulation check to determine if people were making the right kind of judgment (should vs. would), and the instructions did not specify what was meant by "should". Nevertheless, we report the methods and results here for transparency.

This study was preregistered on As Predicted (#28776; anonymous link for peer review: https://aspredicted.org/blind.php?x=ve3ky9).

2. Method

2.1. Participants

Three hundred and ninety-nine native English-speakers in the United States were recruited through Amazon Mechanical Turk (59% women, 40% men, 0.3% other; M_{age} = 38.87, SD = 12.95). The self-reported racial/ethnic makeup of participants was: 6% Asian, 10% Black or African American, 5% Hispanic or Latino, 1% Native American, 0.5% Native Hawaiian or Pacific Islander, 78% white, and 1% other.

Participants were excluded for providing the same name more than once (N = 13), for being non-native English speakers (N = 2), and for saying that their data were not valid (N = 2). Twenty more were excluded due to an experimenter error in the survey software which caused some participants to see the same dilemma more than once, and interfered with recording their response. This left us with a final sample of N = 362.

2.2. Procedure

As in the four studies in the main text, participants each responded to eight dilemmas. For each, participants were asked either what they should, or would, do. Judgment was a blocked within-subjects factor: participants were asked what they would do for the first four dilemmas, and what they should do for the second four, or the reverse. Additionally, participants were either asked what they

thought "you, personally" or "people, generally" would/should do. Judgment target was constant across all trials for each participant. Study S1 thus used a 4-way design: 2 (relationship: close, distant; within-subjects) x 2 (judgment: should, would; within-subjects) x 2 (block order: should first, would first; between-subjects) x 2 (target: you, people; between-subjects).

These methods lacked two key elements that were included in later studies. First, the instructions did not include the clarifying explanations of each judgment type discussed in the main text studies—participants were simply asked what they "would actually" or "should ideally" do. Second, there was no manipulation check for judgment type.

3. Results

Likelihood of reporting was reverse-coded; higher scores thus indicate greater likelihood of protecting the transgressor. We fit a linear mixed model, including participant and dilemma as random intercepts and relationship and judgment type as random slopes on dilemma, the maximal model that reached convergence.

First, we found a main effect of judgment target ("you personally" vs. "people, generally"), showing that people overall reported they themselves both would and should protect more (M = 2.85, SD = 1.88) than people in general (M = 2.49, SD = 1.65; b = 0.36, t(358.01) = 2.73, p < .01, 95% CI = [0.10, 0.61]). Target was also involved in a significant four-way interaction (b = -1.00, t(1802.02) = -3.91, p < .001, 95% CI = [-1.51, -0.49]). However, as this was an exploratory factor and we had no specific predictions involving a four-way interaction, we collapsed across judgment target for the remaining analyses and will not discuss this factor further.

There was a significant three-way Relationship x Judgment x Block Order interaction (b = 0.90, t(1801.38) = 7.03, p < .001, 95% CI = [0.65, 1.15]), shown in Figure E.1. To interpret this interaction, we ran ten post-hoc pairwise contrasts to interpret this interaction, using a Bonferroni correction, giving us a corrected p-critical value of 0.005. For participants who made "would" judgments first, the results looked very similar to those reported in main text Studies 1a and 1b. Participants said they both *would* and *should* protect close others (would: M = 3.77, SD = 1.86; should: M = 2.59, SD = 1.77; b = 1.44, t(539.91) = 13.79, p < .001, 95% CI = [1.24, 1.64]) more than distant others (would: M = 2.34, SD = 1.56; should: M = 1.84, SD = 1.32; b = 0.75, t(539.77) = 7.21, p < .001, 95%CI = [0.55, 0.95). Would-first participants also said that they would protect close others more than they should (b = -1.19, t(589.98) = -12.55, p < .001, 95% CI = [-1.37, -1.01]); the same pattern emerged for distant others, although the would/should discrepancy was smaller (b = -0.50, t(590.40) = -5.32, p < .001, 95% CI = [-0.68, -0.32]).

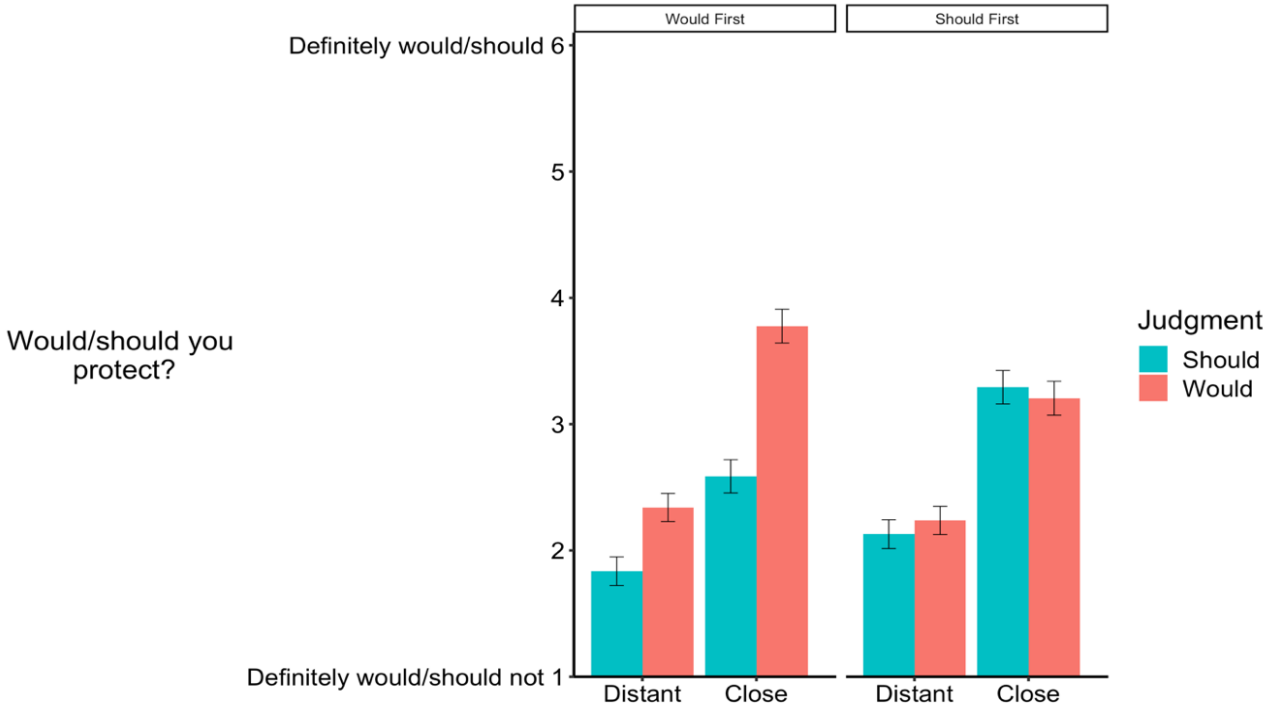
In contrast, participants who made "should" judgments first did not reveal the same discrepancy between judgment types. These participants indicated overall more protecting for close others (*would*: $M = 3.21$, $SD = 1.86$; *should:* $M = 3.29$, $SD = 1.90$) than distant others (*would*: $M = 2.24$, $SD = 1.45$; *should*: $M = 2.13$, $SD = 1.51$; *would:* $b = 0.97$, $t(540.22) = 9.24$, $p < .001$, 95% CI = [0.77, 1.17]; *should:* $b = 1.16$, $t(540.10) = 11.11$, $p < .001$, 95% CI = [0.96, 1.36]), but there was no difference between "should" and "would" judgments for either close ($b = -0.11$, $t(590.01) = -1.14$, $p = .36$, 95% CI = [-0.31, 0.09]) or distant others ($b = 0.09$, $t(590.88) = 0.91$, $p = 0.36$, 95% CI = [-0.11, 0.27]). In other words, when participants were asked to make "should" judgments first, the effect of judgment disappeared – there was no difference between what people said they would and should do.

To further understand the effect of judgment order, we tested how the protecting responses of participants in the should-first condition—where there was no difference between what people said they would and should do—compared to the protecting responses of participants in the would-first condition.

Should-first participants' "would" and "should" protecting responses were significantly higher than what would-first participants said they *should* do ($b = 0.71$, $t(384.62) = 3.99$, $p < .001$, 95% CI = [0.38, 1.04]), *and* were significantly lower than what would-first participants said they *would* do ($b = -0.57$, $t(383.82) = -3.18$, $p < .01$, 95% CI = [-0.92, -0.22]). In other words, should-first participants' judgments about protecting close others fell in between what would-first participants thought they should and would do regarding close others.

Figure E.1

Study S1: Relationship and Judgment Effects across Block Order



Note. Error bars show standard error of the mean.

4. Discussion

Participants who made "would" judgments first showed the same pattern of responses observed in all studies reported in the main text: participants said they both would and should protect close others more than distant others, but also that they *would* protect more than they *should*, especially for close others.

In contrast, no difference between "should" and "would" judgments emerged for participants who made "should" judgments first. However, these results are very difficult to interpret given the crucial methodological limitations noted previously: there was no manipulation check for judgment type, and the instructions lacked clarifying explanations of judgments. In the absence of these elements, we cannot be sure that the participants who were first asked what they should do interpreted this clearly as a normative moral question, rather than a pragmatic and pseudo-predictive one. This worry is supported by data from the manipulation check failures of later studies: across all studies, far more participants failed the manipulation check by selecting "would" when they were supposed to select "should" (e.g. 85% of manipulation check failures in Study 1a, 78% in 1b), than any other kind of failure.

5. Supplemental Study 2

In Supplemental Study 2 (S2), we replicated the blocked-judgment structure of S1, incorporating improved methods that included clearer judgment instructions and judgment manipulation checks (used in Studies 1a and 1b). As in all main-text studies, all questions were about what "you should/would do;" we eliminated the judgment target factor from S1. Because Study 2 presents a stronger test of the effects of making both kinds of judgments, we present that study in the main text; however, we report the present findings for transparency.

This study was preregistered with As Predicted (#35667; anonymous link for peer review: https://aspredicted.org/blind.php?x=3qk4kz).

6. Method

6.1. Participants

Four hundred and three native English-speakers in the United States were recruited through Amazon Mechanical Turk (54% women, 46% men, 0.5% other; M_{age} = 37.94, SD_{age} = 12.42). The self-reported racial/ethnic makeup of participants was (where participants could select as many options as applied to them): 6% Asian, 10% Black or African American, 4% Hispanic or Latino, 1% Native American, 0.3% Native Hawaiian or Pacific Islander, 79% white, and 2% other.

Participants were excluded for the following criteria: providing the same name more than once (N = 22); saying they were not native English speakers (N = 2); saying their data were not valid (N = 7); and failing either of the manipulation check questions (N = 73; all pre-registered criteria). This gave us a final sample N = 299. Results did not differ when we included participants who failed the manipulation check.

6.2. Procedure

We used the same basic design as in Study S1, in which participants responded to eight punish-or-protect dilemmas. In the first four dilemmas they were asked what they would do, and in the second four they were asked what they should do, or the reverse (i.e. order was counterbalanced across participants).

We made the following changes in S2 from the initial S1 design. First, we eliminated the judgment target factor, and (as in Studies 1a and 1b) simply asked participants "what you would/should." We also added the clarified instructions from Study 1b: we explicitly contrasted the question of what one would do ("how you would actually behave in the real world") with what one should do ("the ideal, right thing to do"). Finally, as in 1a and 1b, we added a manipulation check: after each judgment block, we asked participants what kind of judgment they were supposed to be

making in the preceding dilemmas. Thus, S2 has used a 2 (relationship: close vs. distant; within-subjects) x 2 (judgment: should vs. would; within-subjects blocked) x 2 (order: should first vs. would first; between-subjects) design.

7. Results

Likelihood of reporting was reverse-coded; higher scores here indicate greater likelihood of protecting the transgressor. We used a linear mixed model, including participant and dilemma as random intercepts, the maximal model that reached convergence. We used a Tukey correction for follow-up tests to interpret interactions (Tukey-corrected *p*-values reported for simple effects).

A significant Relationship x Judgment Type x Block interaction emerged (b = 0.75, t(2080.55) = 4.01, p < .001, 95% CI = [0.38,1.12]); findings displayed in Figure E.2. To further explore the effects of judgment and relationship, we broke down the data by block order and ran two separate two-way models.

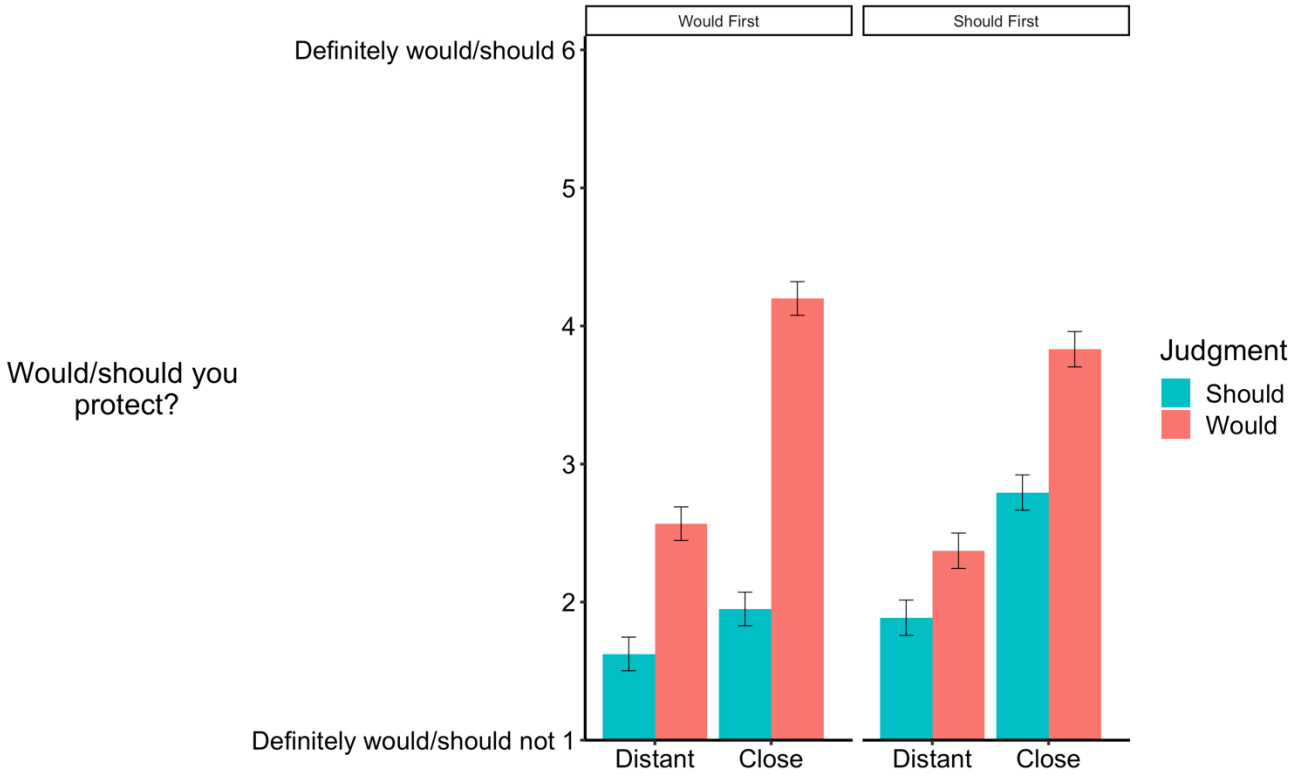
We first looked at participants who made "would" judgments first. As in S1, would-first participants indicated that they would protect close others more than distant others (b = 1.63, t(2080.61) = 17.89, p < .001, 95% CI = [1.45, 1.81). However, unlike in S1 and our other studies, participants did *not* say that they *should* protect close others more than distant others (b = 0.33, t(2080.09) = 3.56, p > .05, 95% CI = [0.15, 0.51]). Finally, would-first participants said they *would* protect more than they *should*, for both close (b = -2.25, t(2080.52) = -24.67, p < .001, 95% CI = [-2.43, -2.07]) and distant others (b = -0.94, t(2080.34) = -10.35, p < .001, 95% CI = [-1.12, -0.76]).

Next we examined participants who made "should" judgments first. These should-first participants said that they would protect close others more than distant others (b = 1.46, t(2080.48) = 15.13, p < .001, 95% CI = [1.26, 1.66]), and also that they should do this (b = 0.91, t(2080.28) = 9.40, p < .001, 95% CI = [0.71, 1.11]). Additionally, they responded that they both *would* and *should* protect close others more than distant others. Additionally, the discrepancy between judgment types emerged for both close others (b = -1.04, t(2080.77) = -10.76, p < .001, 95% C = [-1.24, -0.84]) and distant others (b = -0.49, t(2080.36) = -5.03, p < .05, 95% CI = [-0.69, -0.29]).

The three-way interaction appeared to be driven by an order effect for "should" judgments about close others. Should-first participants responded that they should protect close others more than would-first participants did (b = 0.84, t(506.81) = 5.08, p < .01, 95% CI = [0.51, 1.17]). No other pairwise comparisons across order were significant after correcting for multiple comparisons.

Figure E.2

Study S2: Relationship and Judgment across Block Order



Note. Error bars show standard error of the mean.

8. Discussion

The results in this study were consistent with the overall pattern reported in our main-text studies. Participants in both block order conditions reported that they *would* protect close (and distant) others more than they *should* protect them. Participants who made "should" judgments first said that they both would *and should* protect close others more than distant others, consistent with the pattern reported in our main-text studies. Participants who made "would" judgments first did *not* report that they should protect close others more than distant others. This order block is the only instance across all studies presented here where there was not a significant difference between what people thought they should do regarding close and distant others.

These two supplements overall present further evidence for both the Moral Partiality Hypothesis, and for a reliable discrepancy between what people think they would and should do upon witnessing the moral transgression of a close other.

APPENDIX F

Chapter 4 Moral Transgressions

Below is the complete list of transgressions used in the study.

High Severity:

1. Breaking into the principal's locked office and taking their wallet.
2. Stealing the math teacher's laptop computer out of the locked desk drawer.
3. Taking money out of another student's backpack.
4. Breaking the lock on another student's bike and taking it.

Low Severity:

1. Taking two cookies from the cafeteria, but only paying for one.
2. Working at a bake sale fundraiser and eating some brownies on sale.
3. Taking gum from another student's desk.
4. Sneaking into a high school football game without paying.

APPENDIX G

Chapter 4 Supplemental Undergraduate Study

1. Study Introduction

In this supplemental study, we sought to test whether the well-documented relationship by severity interaction observed in previous adult studies (Berg et al., 2021; Weidman et al., 2020) would replicate in the school context employed in our adolescent study. To do this, we ran the same experiment reported in the main text with current college students over 18 years old: a sample of young adults who are still immersed in a school context.

We preregistered competing predictions regarding the interaction: on the one hand, prior adult work has shown a replicable interaction effect in which adults report close others less than distant others particularly for high-severity transgressions, across transgression contexts—this motivates the prediction that we would see this same interaction with this sample. On the other hand, our adolescent data—which were collected using school rather than "real-world" scenarios used in all prior research with adults—did not reveal an interaction, motivating the alternative prediction that we would not see the interaction effect in school contexts with young adults. This study was preregistered on the Open Science Framework; blinded preregistration for peer review available in "Files" at https://osf.io/znstf/?view_only=02c0894e02aa416aad618e8493310a4e.

2. Method

2.1. Participants

We recruited 961 current undergraduate students in the United States from Amazon Mechanical Turk (35%) and Prolific (65%). We screened for participants who were not on multiple crowdsourcing platforms to avoid duplicate respondents. (We preregistered that all data would be collected from Mechanical Turk, but transitioned to Prolific when we had trouble recruiting enough participants to fit our narrow criteria.) Participants were excluded from analysis for the following criteria: saying they did not answer honestly (n = 10), saying they were not currently in college or listing their grade as a graduate degree (n = 37), or having a duplicate IP address (n = 1; only duplicate response excluded), for a final undergraduate sample of N = 913.

The final undergraduate sample consisted of 54% women, 44% men, and 2% non-binary/other, with a mean age of 22.46 (*SD* = 5.22). The sample was 23% Asian, 11% Black/African American, 15% Hispanic/Latino, 1% Native Hawaiian/Pacific Islander, 1% Native American, 57% white, and 1% other.

Participant information—including exclusion numbers and demographics of the adolescent sample—are reported in the main text. The full dataset consisting of the data from the main experiment (reported in the main text) and this additional undergraduate sample contained *N* = 1826.

2.2. Procedure

As in the study reported in the main text, participants were first asked to imagine their best friend at school, and another student from their school whom they do not know very well. They were asked to think of a specific person and hold that person in mind for the rest of the survey. They were then presented with eight vignettes, in which they imagined witnessing either their best friend, or the distant student, committing either a high- or low-severity theft. Participants were then asked to imagine a professor approaching them and asking whether they had seen anything suspicious, and indicated on a 6-point Likert scale whether they would report the transgressor (1 = "Definitely would not report;" 6 = "Definitely would report"). Participants rated their level of overall trust in professors (0 = "Very untrustworthy"; 6 = "Very trustworthy") as an exploratory measure, and completed a set of demographic questions.

We used the same stimuli as with the adolescent sample, with minor wording adjustments where appropriate for making the school context relevant to undergraduates (e.g., sneaking into a "college soccer game" instead of "high school football game"; authority figure was a professor instead of a teacher).

3. Results

As preregistered, we ran a multilevel model using the lme4 package in R (Bates et al., 2015), combining our adolescent data (collected in the main text study) and the undergraduate data collected here, allowing us to directly test for age group effects. We included participant and dilemma as random intercepts; for random slopes, severity and relationship were allowed to vary by participant, and relationship was allowed to vary by dilemma. A likelihood-ratio tested indicated that this maximal model including these random slopes provided a better fit for the data than a model that only included random intercepts ($\chi^2(7) = 1198.1$, $p < .001$).

Overall, participants were less likely to report close others (*M* = 3.33, *SD* = 1.84) than distant others (*M* = 3.97, *SD* = 1.84; *b* = -0.64, *t*(6.81) = -12.38, *p* < .001, 95% CI = [-0.76, -0.52]), and less

likely to report for low severity (M = 2.89, SD = 1.74) than high severity transgressions (M = 4.41, SD = 1.68; b =1.52, t(6.57) = 14.48, p < .001, 95% CI = [1.28, 1.76]). Undergraduates (M = 3.33, SD = 1.87) were less likely overall to report than were adolescents (M = 3.97, SD = 1.82; b = -0.64, t(1823.79) = 11.66, p < .001, 95% CI = [-0.74, -0.54]). Table G.1 displays the full model fixed effects statistics. These main effects were qualified by a significant three-way interaction among relationship, severity, and age group (b = -0.58, t(9089.44) = -8.55, p < .001, 95% CI = [-0.72, -0.44]). As preregistered, we thus examined the effects of relationship and severity separately by age group. Full results for the adolescent sample are reported in the main text; below we report only the results for the undergraduate sample.

Table G.1

Full Three-Way Model Statistics

Effect	Estimate	*SE*	95% CI		*p*
			LL	*UL*	
Relationship [a]	-0.64	.05	-0.52	-0.76	<.001
Severity [b]	1.52	.10	1.28	1.76	<.001
Age Group [c]	-0.63	.05	-0.74	-0.53	<.001
Relationship x Severity	-0.21	.10	-0.46	0.04	.08
Relationship x Age Group	-0.21	.05	-0.31	-0.11	<.001
Severity x Age Group	0.91	.06	0.79	1.03	<.001
Relationship x Severity x Age Group	-0.58	.07	-0.72	-0.44	<.001

Note. N = 1826. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] -0.5 = distant, 0.5 = close. [b] -0.5 = low, 0.5 = high. [c] -0.5 = adolescents, 0.5 = undergraduates.

Undergraduates

The maximal model again failed to converge with only undergraduate data, so we removed the random slope of relationship varied across dilemma.

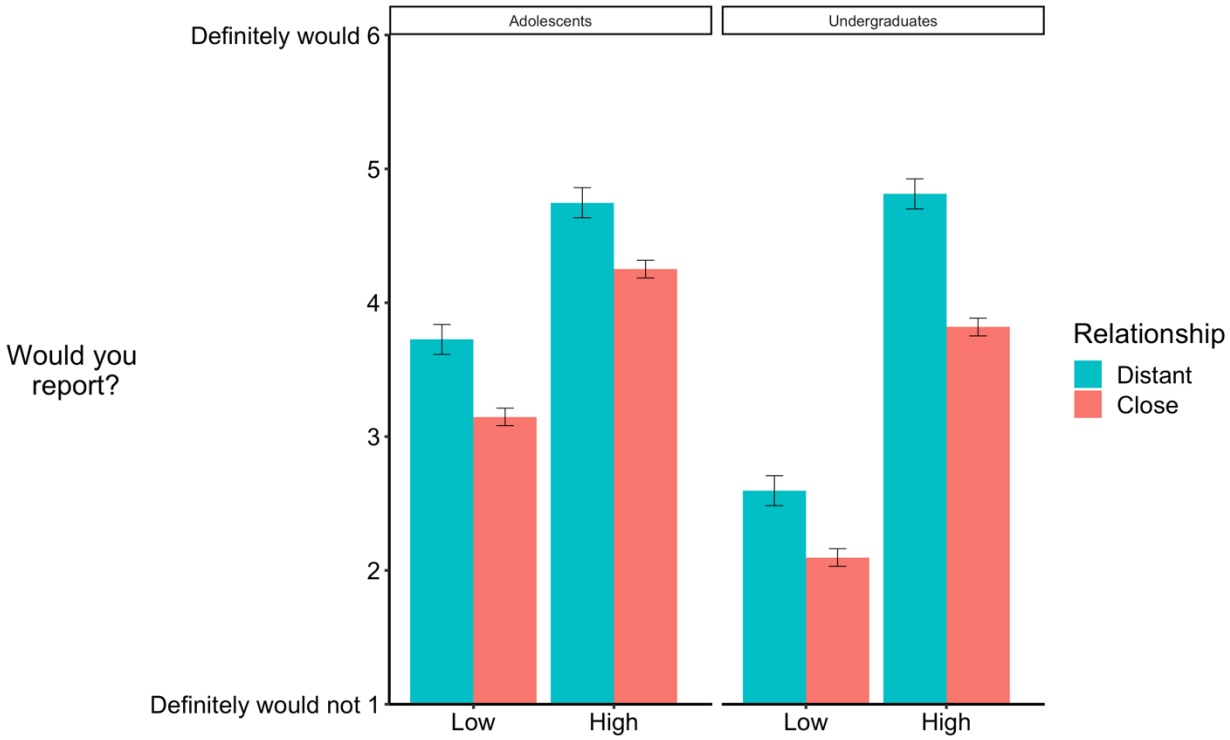
Undergraduates were less likely to report close others (M = 2.96, SD = 1.78) than distant others (M = 3,71, SD = 1.88; b = -0.75, t(1815.24) = -22.09, p < .001, 95% CI = [-0.81, -0.69]), and less likely to report low severity transgressions (M = 2.35, SD = 1.51) than high severity ones (M = 4.32, SD = 1.66; b = 1.97, t(7.55) = 18.11, p < .001, 95% CI = [1.71, 2.23]). However—unlike in our adolescent sample reported in the main text—there was also a significant relationship by severity interaction (b = -0.50, t(9084.47) = -10.36, p < .001, 95% CI = [-0.60, -0.40]). Planned comparisons revealed that undergraduates reported close others less than distant others for both high severity (b = -0.99, t(3889.63) = -24.02, p < .001, 95% CI = [-1.07, -0.91]) and low severity transgressions (b = -

0.50, $t(3889.74) = -12.06$, $p < .001$, 95% CI = [-0.58, -0.42]), but the difference was greater for high severity. Results for adolescents and undergraduates are displayed in Figure G.1.

An exploratory model controlled for professor trust ($M = 3.55$, $SD = 0.84$); higher trust predicted higher reporting ($b = 0.28$, $t(1811.56) = 6.51$, $p < .001$, 95% CI = [0.20, 0.36]), but did not change the overall pattern of relationship and severity effects. We also tested for differences between recruitment platforms. Participants from Prolific were overall less likely to report than those from Mechanical Turk ($b = -0.18$, $t(910.99) = -2.40$, $p < .05$, 95% CI = [-0.33, -0.03]), and there was a severity by platform interaction ($b = 0.43$, $t(910.95) = 4.79$, $p < .001$, 95% CI = [0.25, 0.61]). Despite these effects, the main patterns of results (main effects of relationship and severity, and a relationship by severity interaction) remained constant when we controlled for platform.

Figure G.1

Reporting in Adolescents and Undergraduates



Note. Error bars show standard error of the mean.

4. Discussion

The goal of this study was to test whether the interaction between severity and relationship observed in prior adult work (e.g., Berg et al., 2021; Weidman et al., 2020) would replicate in the school context. We did see a replication of this effect: undergraduate participants were less likely to report

close others than distant others—and the reluctance to report close others was especially pronounced for high-severity transgressions.

This provides evidence that the lack of interaction observed in our early adolescent sample may be indicative of a developmental difference: as children progress into early adulthood, they continue to become overall less likely to report, and the bias against reporting close others begins to become most dramatic when the transgression is serious. It is noteworthy, however, that the racial and socioeconomic demographics of samples varied in key ways. The undergraduate sample was much more heavily white and higher in socioeconomic status than our early adolescent sample. Thus, future research is needed to examine the factors that may contribute to differences across the samples that we have documented.

REFERENCES

Alicke, M., Gordon, E., & Rose, D. (2013). Hypocrisy: What counts? *Philosophical Psychology*, *26*(5), 673–701. https://doi.org/10.1080/09515089.2012.677397

Alston, W. P. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives, 2,* 257–99. https://doi.org/10.2307/2214077.

Archard, D. (1995). Moral partiality. *Midwest Studies In Philosophy*, *20*(1), 129–141. https://doi.org/10.1111/j.1475-4975.1995.tb00308.x

Aristotle. (2009). *The Nicomachean Ethics* (L. Brown, Ed.; D. Ross, Trans.). Oxford University Press.

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality and Social Psychology*, *60*(2), 241–253. https://doi.org/10.1037/0022-3514.60.2.241

Arpaly, N, and Brinkerhoff, A. (2018). Why epistemic partiality is overrated. *Philosophical Topics, 46* (1), 37–51. https://doi.org/10.5840/philtopics20184613.

Bak, I. M. den, & Ross, H. S. (1996). I'm telling! The content, context, and consequences of children's tattling on their siblings. *Social Development*, *5*(3), 292–309. https://doi.org/10.1111/j.1467-9507.1996.tb00087.x

Barbosa, S., & Jimenez Leal, W. (2017). It's not right but it's permitted: Wording effects in moral judgement. *Judgment and Decision Making*, *12*, 308–313.

Baron, M. (1991). Impartiality and friendship. *Ethics*, *101*(4), 836–857. https://doi.org/10.1086/293346

Bartel, C. (2019). Hypocrisy as either deception or akrasia. *The Philosophical Forum*, *50*(2), 269–281. https://doi.org/10.1111/phil.12220

Basu, R. (2018). Can beliefs wrong? *Philosophical Topics. 46*(1), 1–18.

Basu, R. (2019a). Radical moral encroachment: The moral stakes of racist beliefs. *Philosophical Issues*, *29*(1), 9–23. https://doi.org/10.1111/phis.12137

Basu, R. (2019b). What we epistemically owe to each other. *Philosophical Studies*, *176*(4), 915–931. https://doi.org/10.1007/s11098-018-1219-z

Basu, R. (2019c). The wrongs of racist beliefs. *Philosophical Studies*, *176*(9), 2497–2515. https://doi.org/10.1007/s11098-018-1137-0

Basu, R., & Schroeder, M. (2019). Doxastic wronging. In B. Kim & M. McGrath (Ed.), *Pragmatic Encroachment in Epistemology* (pp. 181–205). Routledge.

Bates, D., Maechler, M., Bolker, B., & Walke, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology*, *68*(6), 1042–1054. https://doi.org/10.1037/0022-3514.68.6.1042

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, *77*(3), 525–537. https://doi.org/10.1037/0022-3514.77.3.525

Begby, E. (2013). The epistemology of prejudice. *Thought: A Journal of Philosophy*, *2*(2), 90–99. https://doi.org/10.1002/tht3.71.

Berg, M. K., Kitayama, S., & Kross, E. (2021). How relationships bias moral reasoning: Neural and self-report evidence. *Journal of Experimental Social Psychology*, *95*, 104156. https://doi.org/10.1016/j.jesp.2021.104156

Berndt, T. J. (1982). The features and effects of friendship in early adolescence. *Child Development*, *53*(6), 1447–1460. https://doi.org/10.2307/1130071

Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair": Eight-year-olds reject two forms of inequity. *Cognition*, *120*(2), 215–224. https://doi.org/10.1016/j.cognition.2011.04.006

Bolinger, R. J. (2020). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, *197*(6), 2415-2431.

Bolinger, R. J. (2020). Varieties of moral encroachment. *Philosophical Perspectives, 34*(1), 5–26. https://doi.org/10.1111/phpe.12124.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-tyle moral dilemmas. *Psychological Science*, *29*(7), 1084–1093. https://doi.org/10.1177/0956797617752640

Brank, E. M., Woolard, J. L., Brown, V. E., Fondacaro, M., Luescher, J. L., Chinn, R. G., & Miller, S. A. (2007). Will they tell? Weapons reporting by middle-school youth. *Youth Violence and Juvenile Justice*, *5*(2), 125–146. https://doi.org/10.1177/1541204006296171

Brans, K., Koval, P., Verduyn, P., Lim Y. L., & Kuppens, P. (2013). The regulation of negative and positive affect in daily life. *Emotion, 13*(5), 926–39. https://doi.org/10.1037/a0032400.

Bratman, M. E. (1992). Practical reasoning and acceptance in a context. *Mind*, *101*(401), 1–15.

Brinkerhoff, A. (forthcoming). The cognitive demands of friendship. *Pacific Philosophical Quarterly*.

Brown, B. B., & Larson, J. (2009). Peer relationships in adolescence. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of Adolescent Psychology* (p. adlpsy002004). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470479193.adlpsy002004

Bukowski, W. M., Hoza, B., & Boivin, M. (1993). Popularity, friendship, and emotional adjustment during early adolescence. *New Directions for Child and Adolescent Development*, *1993*(60), 23–37. https://doi.org/10.1002/cd.23219936004

Butler, E. A., Egloff, B., Wlhelm, F. H., Smith, N.C., Erickson, E. A., & Gross, J. J. (2003). The social consequences of expressive suppression. *Emotion*, *3*(1), 48–67. https://doi.org/10.1037/1528-3542.3.1.48.

Bermúdez, J. P. (2021). The skill of self-control. *Synthese*, *199*(3), 6251-6273.

Campbell-Sills, L., Barlow, D.H., Brown, T.A., & Hofmann, S.G. (2006). Acceptability and suppression of negative emotion in anxiety and mood disorders. *Emotion* ,*6*(4), 587–95. https://doi.org/10.1037/1528-3542.6.4.587.

Chiu Loke, I., Heyman, G. D., Forgie, J., McCarthy, A., & Lee, K. (2011). Children's moral evaluations of reporting the transgressions of peers: Age differences in evaluations of tattling. *Developmental Psychology*, *47*(6), 1757–1762. https://doi.org/10.1037/a0025357

Chiu Loke, I., Heyman, G. D., Itakura, S., Toriyama, R., & Lee, K. (2014). Japanese and American children's moral evaluations of reporting on transgressions. *Developmental Psychology*, *50*(5), 1520–1531. https://doi.org/10.1037/a0035993

Cipolletti, H., McFarlane, S., & Weissglass, C. (2016). The moral foreign languge effect. *Philosophical Psychology*, *29*(1), 23–40. https://doi.org/10.1080/09515089.2014.993063

Cohen, L. J. (1989). Belief and acceptance. *Mind, 98*(391), 367–89.

Cohen, L.J., (1992). *An Essay on Belief and Acceptance*. Clarendon Press.

Confucius. (2005). *The Analects of Confucius* (A. Waley, Trans.). Psychology Press.

Csikszentmihalyi, M. (2020). Confucius. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/confucius/

Costin, S. E., & Jones, D. C. (1992). Friendship as a facilitator of emotional responsiveness and prosocial interventions among young children. *Developmental Psychology*, *28*(5), 941–947. https://doi.org/10.1037/0012-1649.28.5.941

Cusimano, C., & Goodwin, G. P. (2019). Lay beliefs about the controllability of everyday mental states. *Journal of Experimental Psychology: General*, *148*(10), 1701–1732. https://doi.org/10.1037/xge0000547

Cusimano, C., & Lombrozo, T. (2021). Morality justifies motivated reasoning in the folk ethics of belief. *Cognition*, 104513. https://doi.org/10.1016/j.cognition.2020.104513

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796. https://doi.org/10.1038/nature05651

Dunning, D. (2007). Self-image motives and consumer behavior: How sacrosanct self-beliefs sway prefences in the marketplace. *Journal of Consumer Psychology*, *17*(4), 237–249. https://doi.org/10.1016/S1057-7408(07)70033-5

Elenbaas, L., Rizzo, M. T., Cooley, S., & Killen, M. (2016). Rectifying social inequalities in a resource allocation task. *Cognition*, *155*, 176–187. https://doi.org/10.1016/j.cognition.2016.07.002

Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations, 1*(2), 140–51. https://doi.org/10.1080/10001998058538695.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004

Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, *425*(6960), 785–791. https://doi.org/10.1038/nature02043

Feldman, R., & Conee, E. (1985). Evidentialism. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *48*(1), 15–34.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441. https://doi.org/10.1016/j.cognition.2012.02.001

Festinger, L. (1957). *A Theory of Cognitive Dissonance* (Vol. 2). Stanford University Press.

Fleisher, W. (2018). Rational endorsement. *Philosophical Studies, 175*(10), 2649–75. https://doi.org/10.1007/s11098-017-0976-4.

Flowerree, A. K. (2017). Agency of belief and intention. *Synthese, 194*(8), 2763–84. https://doi.org/10.1007/s11229-016-1138-5.

Forsberg, C., Thornberg, R., & Samuelsson, M. (2014). Bystanders to bullying: Fourth- to seventh-grade students' perspectives on their reactions. *Research Papers in Education*, *29*(5), 557–576. https://doi.org/10.1080/02671522.2013.878375

Franchow, E. I., & Suchy, Y. (2017). Expressive suppression depletes executive functioning in older adulthood. *Journal of the International Neuropsychological Society*, *23*(4), 341–51.

Franchow, E I., & Suchy, Y. (2015). Naturally-occurring expressive suppression in daily life depletes executive functioning. *Emotion* ,*15*(1), 78–89. https://doi.org/10.1037/emo0000013.

Francis, K. B., Howard, C., Howard, I. S., Gummerum, M., Ganis, G., Anderson, G., & Terbeck, S. (2016). Virtual morality: Transitioning from moral judgment to moral action? *PLOS ONE*, *11*(10), e0164374. https://doi.org/10.1371/journal.pone.0164374

Frankish, K. (2007). *Mind and Supermind*. Cambridge University Press.

Friman, P. C., Woods, D. W., Freeman, K. A., Gilman, R., Short, M., McGrath, A. M., & Handwerk, M. L. (2004). Relationships between tattling, likeability, and social classification: A preliminary investigation of adolescence in residential care. *Behavior Modification*, *28*(3), 331–348. https://doi.org/10.1177/0145445503258985

Frost-Arnold, K. (2014). The cognitive attitude of rational trust. *Synthese*, *191*(9), 1957–1974. https://doi.org/10.1007/s11229-012-0151-6

Gardiner, G. (manuscript). Rape, alcoholism, and selling sex: Against the new ethics of belief.

Gibbs, J. C. (2019). *Moral Development and Reality: Beyond the Theories of Kohlberg, Hoffman, and Haidt*. Oxford University Press.

Graber, J. A., Brooks-Gunn, J., & Petersen, A. C. (2018). *Transitions Through Adolescence: Interpersonal Domains and Context*. Psychology Press.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gregory, A., & Ripski, M. B. (2008). Adolescent trust in teachers: Implications for behavior in the high school classroom. *School Psychology Review*, *37*(3), 337–353. https://doi.org/10.1080/02796015.2008.12087881

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. https://doi.org/10.1016/j.cognition.2009.02.001

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872

Goldberg, S. C. (2019). Against epistemic partiality in friendship: Value-reflecting reasons. *Philosophical Studies*, *176*(8), 2221–2242. https://doi.org/10.1007/s11098-018-1123-6

Gross, J J. (1998a). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology*, *74*(1), 224–37. https://doi.org/10.1037/0022-3514.74.1.224.

Gross, J.J. (1998b). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, *2*(3), 271–99. https://doi.org/10.1037/1089-2680.2.3.271.

Gross, J. J., & Feldman Barrett, L. (2011). Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review*, *3*(1), 8–16. https://doi.org/10.1177/1754073910380974

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being." *Journal of Personality and Social Psychology, 85*(2), 348–62. https://doi.org/10.1037/0022-3514.85.2.348.

Gross, J. J., & Levenson, R.W. (1993). Emotional suppression: Physiology, self-report, and expressive behavior." *Journal of Personality and Social Psychology, 64*(6), 970–86. https://doi.org/10.1037/0022-3514.64.6.970.

Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap. *Social Cognitive and Affective Neuroscience*, *7*(5), 596–603. https://doi.org/10.1093/scan/nsr035

Gyurak, A., Goodkind, M.S., Kramer, J. H., Miller, B.L., & Levenson, R.W. (2012). Executive functions and the down-regulation and up-regulation of emotion." *Cognition and Emotion*, *26*(1), 103–18. https://doi.org/10.1080/02699931.2011.557291.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Hart, D., & Carlo, G. (2005). Moral development in adolescence. *Journal of Research on Adolescence*, *15*(3), 223–233. https://doi.org/10.1111/j.1532-7795.2005.00094.x

Hawley, K. (2014). Partiality and prejudice in trusting. *Synthese*, *191*(9), 2029–2045. https://doi.org/10.1007/s11229-012-0129-4

Hendrickx, M. (manuscript). What is difficulty?

Hester, N., & Gray, K. (2020). The moral psychology of raceless, genderless strangers. *Perspectives on Psychological Science*, 174569161988584. https://doi.org/10.1177/1745691619885840

Heyman, G. D., Chiu Loke, I., & Lee, K. (2016). Children spontaneously police adults' transgressions. *Journal of Experimental Child Psychology*, *150*, 155–164. https://doi.org/10.1016/j.jecp.2016.05.012

Hieronymi, P. (2006). Controlling attitudes. *Pacific Philosophical Quarterly, 87*(1), 45–74. https://doi.org/10.1111/j.1468-0114.2006.00247.x.

Hieronymi, P. (2008). Responsibility for believing. *Synthese*,*161*(3), 357–73. https://doi.org/10.1007/s11229-006-9089-x.

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. *Psychological Review*, *54*(9), 319–340.

Hofmann, W., Brandt, M. J., Wisneski, D. C., Rockenbach, B., & Skitka, L. J. (2018). Moral punishment in everyday life. *Personality and Social Psychology Bulletin*, *44*(12), 1697–1711. https://doi.org/10.1177/0146167218775075

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343. https://doi.org/10.1126/science.1251560

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology*, *56*(3), 561–577. https://doi.org/10.1111/bjso.12199

Hunter, F. T., & Youniss, J. (1982). Changes in functions of three relations during adolescence. *Developmental Psychology*, *18*(6), 806–811. https://doi.org/10.1037/0012-1649.18.6.806

Ingram, G. P. D., & Bering, J. M. (2010). Children's tattling: The reporting of everyday norm violations in preschool settings. *Child Development*, *81*(3), 945–957. https://doi.org/10.1111/j.1467-8624.2010.01444.x

Jankowski, K. F., Moore, W. E., Merchant, J. S., Kahn, L. E., & Pfeifer, J. H. (2014). But do you think I'm cool?: Developmental differences in striatal recruitment during direct and reflected social self-evaluations. *Developmental Cognitive Neuroscience*, *8*, 40–54. https://doi.org/10.1016/j.dcn.2014.01.003

John, O. P., & Gross, J. J. (2004). Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. *Journal of Personality*, *72*(6), 1301–1334. https://doi.org/10.1111/j.1467-6494.2004.00298.x

Johnson, E. J., & Goldstein, D. (2003). Do defaults save lives? *Science*, *302*(5649), 1338–1339. https://doi.org/10.1126/science.1091721

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we have hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368. https://doi.org/10.1177/0956797616685771

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin*, *37*(5), 701–713. https://doi.org/10.1177/0146167211400208

Juvonen, J., & Murdock, T. B. (1995). Grade-level differences in the social value of effort: Implications for self-presentation tactics of early adolescents. *Child Development*, *66*(6), 1694–1705. https://doi.org/10.1111/j.1467-8624.1995.tb00959.x

Kant, I. (1785). *Grounding for the Metaphysics of Morals* (J. W. Ellington, Trans.). Hackett Publishing Company.

Kashdan, T. B. & Steger, M.F. (2006). Expanding the topography of social anxiety: An experience-sampling assessment of positive emotions, positive events, and emotion suppression. *Psychological Science*, *17*(2), 120–28. https://doi.org/10.1111/j.1467-9280.2006.01674.x.

Kawall, J. (2013). Friendship and epistemic norms. *Philosophical Studies*, *165*(2), 349–70. https://doi.org/10.1007/s11098-012-9953-0.

Keller, S. (2004). Friendship and belief. *Philosophical Papers, 33(*3), 329–51. https://doi.org/10.1080/05568640409485146.

Keller, S. (2018). Belief for someone else's sake. *Philosophical Topics*, *46*(1), 19–36. JSTOR.

Kelly, T. (2013). Evidence can be permissive. In M. Steup, J. Turri, & E. Sosa (Eds.), *Contemporary Debates in Epistemology* (pp. 298–311). Wiley-Blackwell.

Kenward, B., & Östh, T. (2012). Enactment of third-party punishment by 4-year-olds. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00373

Killen, M., & Turiel, E. (1998). Adolescents' and young adults' evaluations of helping and sacrificing for others. *Journal of Research on Adolescence*, *8*(3), 355–375. https://doi.org/10.1207/s15327795jra0803_4

King James Bible. (2017). Cambridge University Press.

Koole, S. L. (2009). The psychology of emotion regulation: An integrative review. *Cognition and Emotion, 23*(1), 4–41. https://doi.org/10.1080/02699930802619031.

Kraut, R. (2018). Aristotle's ethics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/aristotle-ethics/

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior*, *33*(4), 323–333. https://doi.org/10.1016/j.evolhumbehav.2011.11.002

LaFontana, K. M., & Cillessen, A. H. N. (2010). Developmental changes in the priority of perceived status in childhood and adolescence. S*ocial Development*, *19*(1), 130–147. https://doi.org/10.1111/j.1467-9507.2008.00522.x

Laurent, S. M., & Clark, B. A. M. (2019). What makes hypocrisy? Folk definitions, attitude/behavior combinations, attitude strength, and private/public distinctions. *Basic and Applied Social Psychology*, *41*(2), 104–121. https://doi.org/10.1080/01973533.2018.1556160

Lee, J., & Holyoak, K. J. (2020). "But he's my brother": The impact of family obligation on moral judgments and decisions. *Memory & Cognition*, *48*(1), 158–170. https://doi.org/10.3758/s13421-019-00969-7

Lee, Y., & Warneken, F. (2020). Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment. *Cognition*, *205*, 104374. https://doi.org/10.1016/j.cognition.2020.104374

Lönnqvist, J.-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology*, *55*, 53–62. https://doi.org/10.1016/j.jesp.2014.06.004

Lord, E. (2016). Justifying partiality. *Ethical Theory and Moral Practice*, *19*(3), 569–590.

Lynch, T. R. Robins, C.J., Morse, J.Q., & Krause, E.D. (2001). A mediational model relating affect intensity, emotion inhibition, and psychological distance. *Behavior Therapy, 32*(3), 519–36. https://doi.org/10.1016/S0005-7894(01)80034-4.

Lyon, T. D., Ahern, E. C., Malloy, L. C., & Quas, J. A. (2010). Children's reasoning about disclosing adult transgressions: Effects of maltreatment, child age, and adult identity. *Child Development*, *81*(6), 1714–1728. https://doi.org/10.1111/j.1467-8624.2010.01505.x

Marshall, J., Mermin-Bunnell, K., & Bloom, P. (2020). Developing judgments about peers' obligation to intervene. *Cognition*, *201*, 104215. https://doi.org/10.1016/j.cognition.2020.104215

Marušić, B., & White, S. (2018). How can beliefs wrong?—A Strawsonian epistemology. *Philosophical Topics*, *46*(1), 97–114.

Mason, C. (2021). The epistemic demands of friendship: Friendship as inherently knowledge-involving. *Synthese*, *199*(1), 2439–2455. https://doi.org/10.1007/s11229-020-02892-w

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10. https://doi.org/10.1016/j.cognition.2014.08.013

McKaughan, D J. (2007). Toward a richer vocabulary for epistemic attitudes: Mapping the cognitive landscape. Ph.D., United States -- Indiana: University of Notre Dame. http://search.proquest.com/docview/304819154/abstract/D48CC791876E46A5PQ/1.

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science*, 0956797619900321. https://doi.org/10.1177/0956797619900321

McRae, K. (2016). Cognitive emotion regulation: A review of theory and scientific findings. *Current Opinion in Behavioral Sciences*, Neuroscience of education, 10 (August): 119–24. https://doi.org/10.1016/j.cobeha.2016.06.004.

Mill, J. S. (1895). *Utilitarianism*. Longmans, Green and Company.

Mills, C. M., & Keil, F. C. (2008). Children's developing notions of (im)partiality. *Cognition*, *107*(2), 528–551. https://doi.org/10.1016/j.cognition.2007.11.003

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez & D. K. Lapsley (Eds.), *Personality, Identity, and Character* (pp. 341–354). Cambridge University Press. https://doi.org/10.1017/CBO9780511627125.016

Misch, A., Over, H., & Carpenter, M. (2018). The whistleblower's dilemma in young children: When loyalty trumps other moral concerns. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00250

Moore, S.A., Zoellner, L.A., & Mollenholt, N. (2008). Are expressive suppression and cognitive reappraisal associated with stress-related symptoms? *Behavior Research and Therapy*, *46*(9), 993–1000. https://doi.org/10.1016/j.brat.2008.05.001.

Morton, J. M., & Paul, S.K. (2019). Grit. *Ethics*, *129*(2), 175–203. https://doi.org/10.1086/700029.

Mulvey, K. L., & Killen, M. (2016). Keeping quiet just wouldn't be right: Children's and adolescents' evaluations of challenges to peer relational and physical aggression. *Journal of Youth and Adolescence*, *45*(9), 1824–1835. https://doi.org/10.1007/s10964-016-0437-y

Niermeyer, M. A., Franchow, E.I, & Suchy, Y. (2016). Reported expressive suppression in daily life is associated with slower action planning. *Journal of the International Neuropsychological Society*, *22(*6), 671–81. https://doi.org/10.1017/S1355617716000473.

Nucci, L. P., & Turiel, E. (1978). Social interactions and the development of social concepts in preschool children. *Child Development*, *49*(2), 400–407. JSTOR. https://doi.org/10.2307/1128704

Olson, K. R., Dweck, C. S., Spelke, E. S., & Banaji, M. R. (2011). Children's responses to group-based inequalities: Perpetuation and rectification. *Social Cognition*, *29*(3), 270–287. https://doi.org/10.1521/soco.2011.29.3.270

Olson, K. R., & Spelke, E. S. (2008). Foundations of cooperation in young children. *Cognition*, *108*(1), 222–231. https://doi.org/10.1016/j.cognition.2007.12.003

Panditharatne, S., Chant, L., Sibley, C. G., & Osborne, D. (2021). At the intersection of disadvantage: Socioeconomic status heightens ethnic group differences in trust in the police. *Race and Justice*, *11*(2), 160–182. https://doi.org/10.1177/2153368718796119

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience*, *9*(1), 94–107. https://doi.org/10.1080/17470919.2013.870091

Paul, S. K. & Morton, J. M. (2018). Believing in others. *Philosophical Topics*, *46*(1), 75-95. *https*://doi.org/10.5840/philtopics20184615.

Peets, K., Hodges, E. V. E., Kikas, E., & Salmivalli, C. (2007). Hostile attributions and behavioral strategies in children: Does relationship type matter? *Developmental Psychology*, *43*(4), 889–900. https://doi.org/10.1037/0012-1649.43.4.889

Pletti, C., Lotto, L., Buodo, G., & Sarlo, M. (2017). It's immoral, but I'd do it! Psychopathy traits affect decision-making in sacrificial dilemmas and in everyday moral situations. *British Journal of Psychology*, *108*(2), 351–368. https://doi.org/10.1111/bjop.12205

Railton, P. (2014). Reliance, trust, and belief. *Inquiry*, *57*(1), 122–150. https://doi.org/10.1080/0020174X.2014.858419

Rapstine, M. (2021). Regrettable beliefs. *Philosophical Studies*, *178*(7), 2169–90. https://doi.org/10.1007/s11098-020-01535-7.

Recchia, H., Wainryb, C., & Pasupathi, M. (2013). "Two for flinching": Children's and adolescents' narrative accounts of harming their friends and siblings. *Child Development*, *84*(4), 1459–1474. https://doi.org/10.1111/cdev.12059

Richards, J. M. (2004). The cognitive consequences of concealing feelings. *Current Directions in Psychological Science, 13*(4), 131–34. https://doi.org/10.1111/j.0963-7214.2004.00291.x.

Richards, J. M., & Gross, J. J. (1999). Composure at any cost? The cognitive consequences of emotion suppression. *Personality and Social Psychology Bulletin*, *25*(8), 1033-1044. https://doi.org/10.1177/01461672992511010.

Richards, J. M., & Gross, J. J. (2000). Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of personality and social psychology*, *79*(3), 410. https://doi.org/10.1037/0022-3514.79.3.410.

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology*, *25*(13), 1731–1735. https://doi.org/10.1016/j.cub.2015.05.014

Rizzo, M. T., Cooley, S., Elenbaas, L., & Killen, M. (2018). Young children's inclusion decisions in moral and social–conventional group norm contexts. *Journal of Experimental Child Psychology*, *165*, 19–36. https://doi.org/10.1016/j.jecp.2017.05.006

Robbins, E., & Rochat, P. (2011). Emerging signs of strong reciprocity in human ontogeny. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00353

Rosenbaum, J. E. (2009). Patient Teenagers? A comparison of the sexual behavior of virginity Pledgers and matched nonpledgers. *Pediatrics*, *123*(1), e110–e120. https://doi.org/10.1542/peds.2008-0407

Ross, H. S., & Bak-Lammers, I. M. D. (1998). Consistency and change in children's tattling on their siblings: Children's perspectives on the moral rules and procedures of family life. *Social Development*, *7*(3), 275–300. https://doi.org/10.1111/1467-9507.00068

Rubin, K. H., Dwyer, K. M., Booth-LaForce, C., Kim, A. H., Burgess, K. B., & Rose-Krasnor, L. (2004). Attachment, friendship, and psychological functioning in early adolescence. *The Journal of Early Adolescence*, *24*(4), 326–356. https://doi.org/10.1177/0272431604268530

Saint-Croix, C. (forthcoming). Rumination and wronging: The role of attention in epistemic morality. *Episteme*.

Scheffler, S. (2010). *Partiality and Impartiality*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199579952.001.0001

Schein, C. (2020). The importance of context in moral judgments. *Perspectives on Psychological Science*, *15*(2), 207–215. https://doi.org/10.1177/1745691620904083

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Schoenfield, M. (2014). Permission to believe: Why permissivism Is true and what it tells us about irrelevant influences on belief. *Noûs*, *48*(2), 193–218. https://doi.org/10.1111/nous.12006

Sebastian, C., Burnett, S., & Blakemore, S.-J. (2008). Development of the self-concept during adolescence. *Trends in Cognitive Sciences*, *12*(11), 441–446. https://doi.org/10.1016/j.tics.2008.07.008

Shah, N. (2002). Clearing space for doxastic voluntarism. *The Monist*, *85*(3), 436-445.. https://doi.org/10.5840/monist200285326.

Shah, N. (2003). How truth governs belief. *The philosophical review*, *112*(4), 447-482.

Shaw, A., & Olson, K. (2011). Children discard a resource to avoid inequity. *Journal of Experimental Psychology. General*, *141*, 382–395. https://doi.org/10.1037/a0025907

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211. https://doi.org/10.1016/j.cognition.2017.05.004

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour*, *2*(10), 750–756. https://doi.org/10.1038/s41562-018-0425-1

Simmons, R. G., & Blyth, D. A. (2017). *Moving into Adolescence: The Impact of Pubertal Change and School Context*. Routledge. https://doi.org/10.4324/9781315124841

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, *1*(3), 229–243. JSTOR.

Slomkowski, C. L., & Killen, M. (1992). Young children's conceptions of transgressions with friends and nonfriends. *International Journal of Behavioral Development*, *15*(2), 247–258. https://doi.org/10.1177/016502549201500205

Smetana, J. (2013). Young children's moral and social-conventional understanding. In M. Banaji & S. Gelman (Eds.), *Navigating the Social World: What infants, children, and other species can teach us* (pp. 352–35). Oxford University Press.

Smetana, J. G. (1981). Preschool children's conceptions of moral and social rules. *Child Development*, *52*(4), 1333–1336. JSTOR. https://doi.org/10.2307/1129527

Sood, S., & Forehand, M. (2005). On self-referencing differences in judgment and choice. *Organizational Behavior and Human Decision Processes*, *98*(2), 144–154. https://doi.org/10.1016/j.obhdp.2005.05.005

Smetana, J. G., & Ball, C. L. (2018). Young children's moral judgments, justifications, and emotion attributions in peer relationship contexts. *Child Development*, *89*(6), 2245–2263. https://doi.org/10.1111/cdev.12846

Soter, L. K. (under review). Acceptance and the ethics of belief.

Soter, L. K., Berg, M. K., Gelman, S. A., & Kross, E. (2021). What we would (but shouldn't) do for those we love: Universalism versus partiality in responding to others' moral transgressions. *Cognition*, *217*, 104886. https://doi.org/10.1016/j.cognition.2021.104886

Sperberg, E. D., & Stabb, S. D. (1998). Depression in women as related to anger and mutuality in relationships. *Psychology of women Quarterly*, *22*(2), 223-238. https://doi.org/10.1111/j.1471-6402.1998.tb00152.x.

Sripada, C. (2018). Addiction and fallibility. *The Journal of Philosophy*, *115*(11), 569–587. https://doi.org/10.5840/jphil20181151133

Sripada, C. (2021). The atoms of self-control. *Noûs*, *55*(4), 800–824. https://doi.org/10.1111/nous.12332

Stalnaker, R. (1984). *Inquiry*. Cambridge, Mass.: MIT Press.

Steup, M. (2012). Belief control and intentionality. *Synthese*, *188*(2), 145-163. https://doi.org/10.1007/s11229-011-9919-3.

Stroud, S. (2006). Epistemic partiality in friendship. *Ethics*, *116*(3), 498-524. https://doi.org/10.1086/500337.

Syvertsen, A. K., Flanagan, C. A., & Stout, M. D. (2009). Code of silence: Students' perceptions of school climate and willingness to intervene in a peer's dangerous plan. *Journal of Educational Psychology*, *101*(1), 219–232. https://doi.org/10.1037/a0013246

Tarrant, M., Dazeley, S., & Cottom, T. (2009). Social categorization and empathy for outgroup members. *British Journal of Social Psychology*, *48*(3), 427–446. https://doi.org/10.1348/014466608X373589

Tassy, S., Deruelle, C., Mancini, J., Leistedt, S., & Wicker, B. (2013). High levels of psychopathic traits alters moral choice but not moral judgment. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00229

Tassy, S., Oullier, O., Duclos, Y., Coulon, O., Mancini, J., Deruelle, C., Attarian, S., Felician, O., & Wicker, B. (2012). Disrupting the right prefrontal cortex alters moral judgement. *Social Cognitive and Affective Neuroscience*, *7*(3), 282–288. https://doi.org/10.1093/scan/nsr008

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00250

Teper, R., Inzlicht, M., & Page-Gould, E. (2011). Are we more moral than we think? Exploring the role of affect in moral behavior and moral forecasting. *Psychological Science*, *22*(4), 553–558. https://doi.org/10.1177/0956797611402513

Thornberg, R., Landgren, L., & Wiman, E. (2018). 'It depends': A qualitative study on how adolescent students explain bystander intervention and non-intervention in bullying situations. *School Psychology International*, *39*(4), 400–415. https://doi.org/10.1177/0143034318779225

Thornberg, R., Tenenbaum, L., Varjas, K., Meyers, J., Jungert, T., & Vanegas, G. (2012). Bystander motivation in buyllying incidents: To intervene or not to intervene? *Western Journal of Emergency Medicine*, *13*(3), 247–252. https://doi.org/10.5811/westjem.2012.3.11792

Traldi, O. (2022). Uncoordinated norms of belief. *Australasian Journal of Philosophy*, *0*(0), 1–13. https://doi.org/10.1080/00048402.2022.2030378

Tuma, E., & Livson, N. (1960). Family socioeconomic status and adolescent attitudes to authority. *Child Development*, *31*(2), 387–399. https://doi.org/10.2307/1125912

Turri, J., Rose, D., & Buckwalter, W. (2018). Choosing and refusing: Doxastic voluntarism and folk psychology. *Philosophical Studies*, *175*(10), 2507–2537. https://doi.org/10.1007/s11098-017-0970-x

Uhlmann, E. L., Zhu, L. (Lei), & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology*, *29*(1), 124–130. https://doi.org/10.1348/026151010X532888

Van Fraassen, B. C. (1985). *Images of Science: Essays on Realism and Empiricism*. University of Chicago Press.

Vasey, M. W., Crnic, K. A., & Carter, W. G. (1994). Worry in childhood: A developmental perspective. *Cognitive Therapy and Research*, *18*(6), 529–549. https://doi.org/10.1007/BF02355667

Viki, G. T., Culmer, M. J., Eller, A., & Abrams, D. (2006). Race and willingness to cooperate with the police: The roles of quality of contact, attitudes towards the behaviour and subjective norms. *British Journal of Social Psychology*, *45*(2), 285–302. https://doi.org/10.1348/014466605X49618

Waldrip, A. M., Malcolm, K. T., & Jensen-Campbell, L. A. (2008). With a little help from your friends: The importance of high-quality friendships on early adolescent adjustment. *Social Development*, *17*(4), 832–852. https://doi.org/10.1111/j.1467-9507.2008.00476.x

Watson, A. J., & Valtin, R. (1997). Secrecy in middle childhood. *International Journal of Behavioral Development*, *21*(3), 431–452. https://doi.org/10.1080/016502597384730

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness–loyalty tradeoff. *Journal of Experimental Social Psychology*, *49*(6), 1027–1033. https://doi.org/10.1016/j.jesp.2013.07.002

Weatherson, B. (2008). Deontology and Descartes's demon. *The Journal of Philosophy*, *105*(9), 540-569.

Williams, B. 1973. Deciding to believe. In B. Williams (Ed.) *Problems of the Self*, 136–51. University Press.

Weidman, A. C., Sowden, W. J., Berg, M. K., & Kross, E. (2020). Punish or protect? How close relationship shape responses to moral violations. *Personality and Social Psychology Bulletin*, *46*(5), 693–708. https://doi.org/10.1177/0146167219873485

Williams, B. (1981). *Moral Luck: Philosophical Papers 1973-1980*. Cambridge University Press.

Wolf, S. (1992). Morality and partiality. *Philosophical Perspectives*, *6*, 243–259. JSTOR. https://doi.org/10.2307/2214247

Wright, C., & Weekes, D. (2003). Race and gender in the contestation and resistance of teacher authority and school sanctions: The case of African Caribbean pupils in England. *Comparative Education Review*, *47*(1), 3–20. https://doi.org/10.1086/373962

Yao, V. (2020). Grace and alienation. *Philosophers' Imprint*, *20*(16), 1–18.

Youniss, J., & Haynie, D. L. M. A. (1992). Friendship in adolescence. [Review]. *Journal of Developmental*, *13*(1), 59–66.

Yu, H., Siegel, J. Z., & Crockett, M. J. (2019). Modeling morality in 3-D: Decision-making, judgment, and inference. *Topics in Cognitive Science*, *11*(2), 409–432. https://doi.org/10.1111/tops.12382

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, *149*(1), 182–191. https://doi.org/10.1037/xge0000613