# BAD BELIEVERS: EVIDENCE-RESISTANCE, RATIONAL PERSUASION, AND SOCIAL CHANGE

By

CAROLINA FLORES HENRIQUE

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Elisabeth Camp and Susanna Schellenberg

And approved by

New Brunswick, New Jersey

October 2022

# ABSTRACT OF THE DISSERTATION

## Bad Believers: Evidence-Resistance, Rational Persuasion, and Social Change

by CAROLINA FLORES HENRIQUE

Dissertation Directors: Elisabeth Camp and Susanna Schellenberg

Can a belief be rational despite being resistant to counter-evidence? Why are some beliefs resistant to counter-evidence in the first place? Given the pervasiveness of evidence-resistant beliefs, how can we institute widespread belief change? My dissertation *Bad Believers* develops a general theory of belief and belief revision that addresses these questions.

The theory I argue for is a form of *rationalism* about belief, the view that there is an important connection between belief and rational agency. I develop and defend it by carefully considering pervasive evidence-resistance and strong external influences on how believers interact with evidence. My aim is to establish not only that rationalism is compatible with these facts, but also that the version of rationalism I defend is needed to account for this data.

This project has two parts. The first part is a defense of the rationalist idea that belief is constitutively evidence-responsive. Specifically, I argue that, if an attitude is a belief, then the subject has the capacity to rationally update it in light of relevant evidence. I support this view by arguing that it best accounts for the descriptive and normative roles of belief. I then apply the view to account for clinical delusions, showing how it can illuminate even deep, pathological evidence-resistant beliefs.

In the second part, I develop theoretical tools to help us understand the role of agency in belief maintenance and revision. Against pure structuralist accounts of belief maintenance, I aim to show that we need to consider individual personality and agency to understand resistant social beliefs. I then build on this work by developing the notion of epistemic style. I use this notion to argue that paradigmatic instances of belief updating (in adult humans) express the agent's *epistemic* personality, making room for sophisticated self-regulation.

The account of belief and belief revision that emerges offers new theoretical resources for thinking about epistemic normativity and about the distinctiveness of the mental. At a social and political level, it aims to provide tools for rational persuasion and for exploring the role of individual beliefs in social change. In this way, my dissertation contributes to a social turn in philosophy of mind, mirroring similar developments in epistemology, philosophy of language, and metaphysics.

# ACKNOWLEDGMENTS

## Acknowledgments

Very many people contributed to my writing this dissertation. First and foremost, I want to thank my dissertation committee. This dissertation owes the most to my two advisors, Liz Camp and Susanna Schellenberg. I came to Rutgers largely because, after a couple of half an hour meetings, I was sure I wanted to work with Liz and Susanna. Those first impressions were right: Liz and Susanna have been the best advisors I could wish for.

Talking with Liz Camp about my dissertation was often the highlight of my week, which, I gather, is not how most people feel about meeting with their advisors. Every time, I would leave the call almost jumping from enthusiasm about getting to do philosophy. Liz would always draw connections I had not seen, bring out the big picture, and find the virtues in my half-baked ideas and help me articulate them. She always encouraged me to keep exploring unwieldy philosophical questions in quirky ways. Liz's own work—putting analytic rigour at the service of exploring the messy complexity of real human minds, bringing in work from a wide variety of traditions, and letting one's own personality shine through—is the ultimate model for my own work. Beyond that, Liz has supported me and helped me grow as a full person, not just a philosopher. Her wise advice and emotional generosity have been crucial to me throughout graduate school.

Susanna Schellenberg is directly responsible for the topic of this dissertation and for its timely completion. A few years ago, when we were both in Germany, Susanna generously took two full days of her schedule in Tubingen to talk with me (at the time, a flustered second-year) about my work. At that time, she suggested I continue to work on an ancestral version of chapter 2, and that I focus my attention on evidence-resistance rather than irrationality more generally; and that was the start of the dissertation. This is only one illustration of Susanna's crucial help in guiding my focus and attention, getting me to carve out my ideas and push them into a cohesive form, and taking projects to completion even when I was struggling to believe in them. Susanna taught me how to structure papers, write in clear and direct ways, and compellingly set up my projects. And her determination, sharpness, and directness have made her a model of how to exist in philosophy spaces.

The rest of my committee was similarly helpful. I am one in a very long list of epistemologists who have been fortunate to learn from Ernie Sosa's systematic thinking, exemplary kindness, and the warm and sharp community Ernie has built. I am grateful to Ernie for generously believing in my work while pressing forceful objections to the project of the dissertation. My work benefited a lot from the critical feedback and wide-ranging discussion at his dissertation workshop, which combined a friendly atmosphere with serious intellectual ambition and a willingness to explore big philosophical questions—virtues that I want to take with me.

When I first met Eric Mandelbaum, I thought of him as a philosophical nemesis. Since then, Eric has brought me around to many of his views, and suspicion has transformed to admiration and friendship. Eric's high energy and enthusiasm for philosophy and psychology are electrifying, and his encyclopedic knowledge and skill at systematizing the terrain in empirically-informed philosophy of mind are traits to which I aspire.

Alex Guerrero has helped me begin to articulate social and political dimensions of my work, and I am grateful to him for making space for me to learn about non-Western traditions. More generally, I am extremely grateful for all of Alex's tireless work to make the Rutgers department, and academic philosophy more generally, more supportive and philosophically inclusive. I can only hope to match his tirelessness and courage as a faculty member.

The papers in this dissertation benefited immensely from feedback from a very large number of people. In addition to my dissertation committee, I want to thank people who gave me comments on, or took time to discuss with me, all or part of the dissertation: Austin Baker, D'i Black, César Cabezas, Laura Callahan, Andy Egan, Frankie Egan, Jon

Ellis, Megan Feeney, Danny Forman, August Gorman, Sally Haslanger, Sahar Heydari Fard, Pamela Hieronymi, Daniel Hoek, Adam Gibbons, Samia Hesni, Sofia Jeppsson, Tyler John, Quill Kukla, Kelsey Laity D'Agostino, Amy Levine, Ting-An Lin, Annette Martín, Matt McGrath, James McIntyre, Albert Newen, Thi Nguyen, Jill North, Dee Payton, Hanna Pickard, Paul Pietroski, Jake Quilty-Dunn, Ezra Rubenstein, Susanna Siegel, and Caroline von Klemperer. These papers improved a great deal from the criticism and suggestions of so many insightful philosophers. Special thanks are due here to Verónica Gómez Sánchez, Chris Willard-Kyle, and Elise Woodard, who heroically gave me line-by-line comments on multiple drafts of (most of) my dissertation, and who have been fantastic philosophical interlocutors over the years.

I am also fortunate to have been part of a vibrant philosophical community throughout graduate school. I learned a huge amount from many of my peers at Rutgers, and am grateful for their encouragement, support, and friendship over the years. In particular, I am grateful to D'i Black, Laura Callahan, Megan Feeney, and Will Fleischer for having patiently shown me the ropes in epistemology, as well as for their continually brilliant work; to Austin Baker, Sam Carter, Verónica Gómez Sánchez, Jimmy Goodrich, Nico Kirk-Giannini, Morgan Moyer, Eli Shupe, Peter van Elswyk, Isaac Wilhelhm, for their warm welcome to Rutgers and for supporting me through some very difficult times early in graduate school; and to my cohort members Adam Gibbons, Tyler John, Ting-An Lin, Ezra Rubenstein, and Chris Willard-Kyle for the shared journey. Finally, I am lucky to have learned a lot from the sharp, inventive, and kind philosophers who joined Rutgers after I did. In particular, I want to thank Mia Accommando, Jordan Bridges, Chris Copan, Cal Howland, Zach Kofi, Lauren Lyons, Andrew Rubner, Lauren Richardson, and Caroline von Klemperer.

I also benefited from being part of a larger network of early career philosophers with social interests, which helped me feel like I could belong in academic philosophy (and actively make it a better place). I am grateful for the many people I organized with in

Minorities and Philosophy (MAP), especially Elise Woodard, Angela Sun, Jingyi Wu, César Cabezas, Eric Bayruns Garcia, Annette Martín, and Yarran Hominh. I am also grateful to Samia Hesni and Elli Neufeld for having helped me explore nascent interests in social philosophy early in graduate school.

In addition to learning from my peers, I also learned a lot from graduate classes and conversations with faculty in the New York area. I am especially grateful to Ned Block, Dave Chalmers, Andy Egan, Frankie Egan, Béatrice Longuenesse, Bob Matthews, and Matt McGrath, who, in graduate seminars and conversation, taught me a lot about philosophy of mind an related topics. Looking further back, I want to thank my undergraduate mentors, who are largely to blame for my love of philosophy. My tutor at Oxford, Adrian Moore, deeply shaped my view of philosophy as a humanistic discipline. He taught me from the get-go that philosophy can and should combine critical acumen with kindness, curiosity, and openness, and my generalist tendencies in philosophy owe a lot to having been taught by someone with such wide knowledge and interests. My undergraduate thesis advisor, Anita Avramides, directed my early work on topics close to this dissertation, suggesting to me that I work on delusions and expertly guiding my exploration of the topic. My work owes a huge amount to her Davidsonian leanings and to her emphasis on personal-level aspects of the mental.

I was also fortunate to have wonderful mentors who displayed for me how to do philosophy in authentic and socially engaged ways. Esa Díaz-León, Meena Krishnamurthy, Lisa Miracchi, Thi Nguyen, Nico Orlandi, Quill Kukla, Kate Ritchie, Susanna Siegel, and Christina van Dyke have all been models to me, and I have treasured their practical advice and philosophical insight. I am also grateful to Alejandro Vásquez del Mercado for crucial help in getting my first publication, and to Ema Sullivan-Bissett for additional opportunities to explore topics in philosophy of psychiatry. Finally, I am deeply grateful to Ruth Chang for her encouragement, empathy, tact, and practical help at a crucial juncture in graduate school for me; I am not sure I would have finished my PhD without her support.

I owe a similar debt to my friends throughout graduate school. In addition to people I have already named (or who I will name below), I am grateful to Emma Atherton, Matt Andler, Rowan Bell, Sarale Ben Ashler, Clifford Carr, Tez Clark, Nina DeMeo, Lindsay Ferris, Vera Flocke, Marta Franco, Natasha Frost, Noga Gratvol, Dan Harris, Ben Henshall, Zoe Johnson King, Jenny Judge, Rahul Kulka, Annina Loets, Alice Martin, Charlie McLean, Rose Ryan Flynn, Beatriz Santos, Keyvan Shafiei, Daniel Sharp, and Susan Wu, for different combinations of inspiration, refreshing my take on the world, bringing adventure and chaos to my life, and supporting me through the ups-and-downs of graduate school.

Finally, I am deeply grateful for the friendship and love of people below. (Warning: this is the extra-cheesy bit of these acknowledgments, so read at your own risk!)

Austin Baker showed me many sides of New York I came to fall in love with, was a source of solidarity and clarity on many of my difficult feelings about graduate school. I have learned a lot from their socially-grounded, empirically-rigorous philosophical approach.

Banafsheh Beizaei is one of the wisest and most generous people I know. She has graciously pulled me out of many dramatic life crises (ranging from break-ups to floods), and been a sweet and joyous companion in many excellent New York and Germany adventures.

Martina Botti gave me space to freely indulge in acting out Southern European stereotypes in New York. Her flamboyant sense of humor and creative energy have brought a lot of joy to my life.

Sophie Côte has added a much-needed layer of non-trashy style to many European adventures. I have learned much from her intellectual seriousness and unflinching way of looking at life.

Tyler John was a wonderful roommate for four years. I am grateful for all the times he sincerely laughed at my teenage jokes, listened to litanies of tearful complaints with deep care, and fed me vegan junk food.

Nico Kirk-Giannini's off-color sense of humor and appreciative commentary on my ridiculous stories make me laugh daily. Perhaps surprisingly given his compelling mean-boy demeanour, Nico uniquely had my back even before he knew me well, and his on-point advice and attentive presence have helped me steer my life in fun-loving and authentic ways.

Amy Levine and I first bonded over graduate school admission anxiety. Since then, we have continued to bond over the joys of coming out, running around, and trying to do philosophy that is actually about and for humans. Her persistence, non-judgmental listening, and warmth have kept me grounded.

Lauren Lyons has been an incredible partner in high-energy nightlife antics, and I aspire to (and sometimes envy) her uniquely expansive and adventurous spirit, political acumen, and open-heartedness.

Dee Payton was the best companion in learning about feminist philosophy I could have wished for. Her style, constant exploration, and openness to the dark and inexplicable have enriched my life in deep ways.

Dena Shottenkirk provided me with a real home in which to finish writing my dissertation. I am inspired daily by her curiosity, questioning spirit, and single-minded determination to do her own thing. She and her granddaughter Frankie Simmons never fail to help see the world in brighter, more playful lights.

Daniel Young has taught me a huge amount about how to do philosophy and think about politics in nuanced progressive ways. He has also been a sweet friend, repeatedly teaching me how to accept complicated emotions in clear-eyed and brave ways.

Isaac Wilhelm has been a generous, supportive friend who I have turned to in many difficult moments in full confidence that I would feel loved. I am fortunate to count on his emotional openness and sensitive way of looking at the world.

Elise Woodard has been an incredible long-distance friend. Elise has been my closest work friend, so much so that it sometimes feels like I get to have a second, better brain available. We have co-authored, co-organized MAP, read all of each other's papers, ran a reading group, and jointly explored and developed our interests in epistemology. I cannot imagine graduate school without her. On top of that, I have been fortunate to be in touch more-or-less daily and with full emotional transparency, and to get Elise's clear insight, empathy, and help with goal-setting, organizing, and celebrating life.

Verónica Gómez Sánchez has been a constant inspiration and a real life-saver. Verónica took me under her wing in my first year, and was my central source of support during a brutal beginning to grad school. I have often stayed at her place, or spent many long hours co-working, snacking, having meandering conversations, and feeling fully at home with her. In addition to being a wonderful friend, Verónica is also one of the most insightful, ambitious, brilliant philosophers I know, and I have learned so much from her. I am hugely grateful that we are both ending up in the same corner of the world.

Caroline Bowman has been my closest friend in graduate school. We have shared so much love, as well as the dizzying experience of coming out in New York in our early 20s and exploring the queer scene together. Caroline has patiently helped me find my footing even when I was not easy to be around, and I am grateful for her open-heartedness and generosity. Beyond that, I find myself constantly inspired by her determination, activist leadership, and courage. On a more hedonistic note, I am grateful for all the times we got to party through the night, explore new neighborhoods, bop around to hyper-pop, and ride our bikes in New York, Lisbon, Berlin, Paris, Mexico City, and the US south and Florida.

My family (the Flores: my grandparents, Zé and Isabel, my uncles and aunts (Patrícia, Pedro, Mónica, Ciro), and my cousins (Mateus, Meg, Rita, and Gui)) somehow managed to love and support me at a very long distance and in the face of many erratic life decisions; and, of course, I owe much of who I am to their care and warmth, large personalities, and argumentative spirit. My dad, Domingos Henrique, inspired me to see academia as continual learning, and learning as the best thing one can do. I am lucky to have a little sister, Beatriz Flores, whose intelligence, prodigious memory, direct manner, and social graces I am constantly proud of. And I am especially grateful to my mom, Isabel Flores, who has always believed in me much more than I could, and whose grit, determination, and resilience are something I aspire to.

### Acknowledgment of previous publications

**P1** Flores, Carolina (2021). Delusional evidence-responsiveness. *Synthese* 199 (3-4):6299-6330. Reproduced with permission from Springer Nature.

**P2** Flores, Carolina (forthcoming). Epistemic styles. *Philosophical Topics*.

# TABLE OF CONTENTS

# CHAPTER 1
# INTRODUCTION

## 1.1 Motivating the Project

This dissertation develops a general theory of belief and its connection to rational agency. It defends a version of rationalism, the view that there is a tight connection between belief and rational agency. In doing so, it develops general conceptual resources for describing and evaluating beliefs and belief revision in real-world agents.

Given the central theoretical role of belief in philosophy, theorizing about belief and rational agency has implications for many long-standing philosophical debates. For example, it bears questions about how to fit reasons in a world of causes and questions about the grounds of epistemic normativity. This project also has practical import. We often understand others' social and political behavior in terms of their beliefs, and criticize one another for having bad beliefs—beliefs that appear disconnected from one's rational agency. The project of this dissertation contributes to understanding such beliefs and when such criticisms are warranted.

This topic has special bite in the current political and cultural circumstances. In response to the last decade's global turn toward illiberal, anti-democratic politics, it has become popular to ascribe deep irrationality to ordinary people. On this narrative, people make wrong political decisions because they have bad beliefs. Indeed, this line goes, they make bad decisions because they are irredeemably bad believers—stupid, childish, irrational, ignorant, and either unable to tell fact from fiction or totally uninterested in having true beliefs. This narrative has been used to support attacks on democratic forms of government (Brennan 2016, Achen and Bartels 2016). If ordinary people are irredeemably bad at reasoning, then their political participation will not secure social goods, and we should make collective decisions in a way that bypasses their beliefs. The account of belief I develop in this dissertation provides resources for resisting this narrative.

The dissertation comprises four self-standing papers. Instead of providing a summary of each of the papers, I will here bring together central themes to explain how the view I defend secures a central place for rational agency in belief. The material in this section goes beyond the contents of the papers in the dissertation. It is meant to provide as big-picture reading guide, and to orient the reader's attention to non-obvious connections between the papers.

The label "rationalist," as I am using it, refers to any view of belief on which there is a tight connection between belief and rational agency. Most prominently, rationalist views include views on which having beliefs requires being rational, or on which an attitude counts as a belief only if it is sufficiently rational. For example, D. Davidson 1985, Dennett 1981, Gendler 2008, Helton 2020, McDowell 1998, M. Smith 2003, D. Velleman 2000, all hold rationalist views. Under the common assumption that humans have many beliefs, rationalist views entail that rational agency plays a central role in human cognition.

The version of rationalism I develop has two parts. First, beliefs constitutively involve rational agency in that they are constitutively evidence-responsive (in a sense of "evidence-responsiveness" I will detail below). Second, paradigmatic beliefs in adult humans are agency-involving in a deeper sense: we characteristically regulate our beliefs in ways that express our epistemic personality, and we often do so in active, flexible ways.

I develop this account by considering four challenges for rationalist views: (1) ordinary evidence-resistance; (2) psychiatric cases of deep evidence-resistance; and the role of (3) social structures and (4) contextual factors in explaining belief maintenance and interactions with evidence. The view I develop is directly guided by on-the-ground empirical facts about the difficulty of belief change, helping us make sense of these otherwise puzzling phenomena. In developing this account, I will make the case that rational agency is a feature of ordinary human cognition, even in cases of seemingly deep irrationality.

## 1.2 Belief & Rationality

As a first challenge for rationalism, consider psychiatric cases of extreme evidence-resistance: clinical delusions. Patients with delusions are unmoved by evidence that is in direct conflict with their delusions, often responding to it by offering obvious, and bizarre, confabulations. If delusions are beliefs—as the DSM claims—then how can beliefs necessarily involve rational agency?

One might think that one can address this challenge by claiming that delusions are too strange to count as beliefs. But this will not get to the root of the problem. For evidence-resistance is pervasive (the second challenge). If beliefs involve rational agency, and if rational agency involves rationally updating one's beliefs in light of the evidence, then how come so many beliefs routinely fail to be rationally updated in the face of counter-evidence? Even worse, how come so many beliefs get stronger in the face of counter-evidence?[1]

Partially in response to such issues, recent theorists have suggested that we ought to think of belief in terms of features that have little to do with epistemic rationality or responsiveness to evidence. Indeed, this suggestion is shared among accounts that otherwise profoundly differ. For example, Eric Schwitzgebel 2002's dispositionalist view ties belief to dispositions to behave, think, and feel in belief-stereotypical ways. Aaron Zimmerman 2018's pragmatist view individuates belief in terms of connections to relatively self-controlled action. And Quilty-Dunn and Mandelbaum 2018's representationalist view takes beliefs to be mental representations that can figure in forward-looking inferences. In all these views, beliefs have no significant connection with rationally responding to evidence.

In contrast, in this dissertation, I argue that beliefs are constitutively responsive to evidence, and therefore that believing and rationality are tightly connected. In chapter 2, I develop and defend the *Capacities View of Belief*, on which, if an attitude is a belief, then

[1] See Bortolotti 2009 andStanley 2015 for versions of this kind of challenge.

it is underwritten by the capacity to rationally respond to evidence. No attitude counts as a belief unless it is in the province of evidence-responsiveness capacities.

### 1.2.1 Addressing "What is belief?"

This dissertation, then, offers a partial answer to the question "What is belief?". Because such questions can be taken in many different ways, I want to clarify my methodology. I take answering such "What is $x$?" questions be a matter of *rationalizing self-interpretation* (L. Schroeter and F. Schroeter 2015). The goal of rationalizing self-interpretation is to discover the original meaning associated with a token representation "$x$," what we—the representational tradition that thinks in $x$-terms—have been thinking and talking about all along.[2]

To find out what that phenomenon is, we start by determining what purposes the concept plays in our representational tradition: what do we want the concept of belief for? Following this first step, we refine our substantive understanding of the topic so as to find out which semantic interpretation can meet the relevant interests: Given the point of the concept of belief, which things in the world does the concept of belief pick out?

My defense of the Capacities View of Belief is guided by two central sets of interests at play when it comes to the concept of belief.

One set of interests that the concept of belief serves are *normative-regulative* interests (McGeer 2007a, Zawidzki 2013): we appeal to the concept of belief to criticize and praise

[2] This method is a more general and systematic version of *function-first epistemology* (Craig 1991, Fricker 2007, Hannon 2018), which aims to investigate knowledge by reflecting on the point of having such a concept. Note that it differs from conceptual analysis: it involves making best sense of scientific results and practices of social interaction that involve the concept of belief. not only (or primarily) systematizing linguistic usage.
It differs, also, from conceptual engineering or conceptual ethics (Burgess and Plunkett 2013, Burgess, Cappelen, et al. 2019, Chalmers 2020) projects, which aim to reform our concept of belief (Zimmerman 2018, Schwitzgebel 2021), not to capture our existing concept. That said, my project shares with engineering projects an emphasis on how our goals constrain our concepts. For this reason, if you disagree with me about the goals that our concept of belief *does* serve, or the relative importance of those goals, you can read the project as one of conceptual engineering, i.e., as proposing a new concept of belief that serves specific goals.

one another, and to hold one another accountable.[3] These practices include assessing and regulating beliefs along epistemic lines. We criticize beliefs for being false, irrational, or overly dogmatic, and frequently take this criticism to redound on the agent who has these beliefs.

To put the point differently: adapting Sellars 1956, in characterizing an episode or state as that of believing, we are placing it in the space of reasons, as the kind of thing that is eligible for assessment and regulation along straightforward epistemic lines. As an influential through line in 20th century philosophy of mind and epistemology held, the concept of belief serves to delimit the bounds of the space of reasons (Brandom 1994, D. Davidson 1982, McDowell 1998, Sellars 1956).

I take regulative-normative interests to be central to the concept of belief. First, such interest distinguish belief from other attitudes: when we ascribe other attitudes, such as desires, imaginings, and so on, we are not automatically placing them in the space of reasons. Second, picking out attitudes in the space of reasons serves practically and socially important purposes. When an attitude is in the space of reasons, we can enter in the social game of giving and responding to reasons about the truth of its content. This opens the door to joint deliberation and arriving at a shared view of reality. These are important social practices, and ones which we have an interest in demarcating with our concepts.

Regulative-normative interests make rationalist views of belief appealing. If beliefs are in the space of reasons—if we hold each other accountable for beliefs on the basis of responsiveness to evidence, and criticize people for failing to adjust their beliefs to the evidence, frequently explain why people hold the beliefs they do in terms of the evidence they have, and expect people to change their beliefs in response to evidence—we should expect belief to be tightly connected to rational agency.

[3] It is because I take such interests—and not exclusively descriptive-explanatory interests—seriously that my project does not reduce to the psychofunctionalist (Quilty-Dunn and Mandelbaum 2018) project of finding which natural kind the concept of belief refers to, where natural kinds are understood as kinds that play an important role in scientific explanation and prediction.

However, there is another set of interests at play that must be accommodated: *descriptive-explanatory* interests. We appeal to the concept of belief to describe, predict, and explain intentional action and other forms of intelligent behavior. To the extent that these folk-psychological generalizations or models (Godfrey-Smith 2005) aim at giving us the most accurate explanations and predictions of behavior, we should see them as proto-scientific, aiming to get at a natural kind. The descriptive-explanatory interests that the concept of belief serves make it important to look at our best science of belief to come to understand what beliefs are, much as is generally the case for natural kinds. For this reason, any passable account of belief must do justice to results in the psychology of belief, including the results that pose challenges to rationalist views.

Showing that, indeed, the Capacities View of Belief can do justice to these interests—and not only to regulative-normative interests at play—is the central task of chapter 2. I argue that my view is supported by our best scientific accounts of evidence-resistant beliefs. Resistance to counter-evidence is the result of active processes aimed at restraining rational capacities. Specifically, the unpleasant feeling of cognitive dissonance serves to motivate us to find byzantine ways of accommodating counter-evidence while preserving our cherished beliefs—such as beliefs in our own goodness and that of our social groups. The fact that such an active, effortful process is needed to preserve these beliefs indicates underlying rational capacities that would otherwise kick into gear, forcing us to live with harsh views of ourselves and the world. The upshot is that a version of the Davidsonian view that to have beliefs is to be a rational creature finds vindication in cognitive science. Indeed, it is vindicated by findings that have often been taken to establish that we are deeply irrational.

More generally, by integrating descriptive-explanatory and normative-regulative interests, we can come to understand what beliefs are. In doing so, we can bring together two diverging strands in analytic philosophy: the Quinean idea that mental kinds are natural kinds that figure in our best science of the mind, and the Strawsonian idea that mental concepts (which pick out these kinds) are thoroughly enmeshed in practices of interpersonal regulation.

### 1.2.2 A better rationalist understanding of irrationality

I put the Capacities View of Belief to the test in chapter 3, where I show that it can accommodate the claim that delusions are beliefs. I argue that, while being profoundly evidence-resistant, delusions are nonetheless evidence-responsive in the sense that they involve capacities to rationally respond to evidence. Contrary to popular belief, clinical delusions do not erase our rational capacities. Their evidence-resistance is the result of psychological factors (including desires, unusual perceptual experiences, and cognitive biases) that mask those capacities.

This account of delusions addresses a long-standing disagreement about the nature of delusions. Specifically, it shows that there is room to hold both that belief is constitutively evidence-responsive and that delusions are beliefs. These views have been taken to be inconsistent. This has motivated some theorists to claim that delusions are not beliefs (but imaginings, acceptances, or *sui generis* attitudes; e.g. Currie and Ravenscroft 2002, Egan 2008a, and Frankish 2012). In the opposite direction, it has motivated others to claim that we must give up the claim that belief is constitutively evidence-responsive (e.g. Bortolotti 2009 and T. J. Bayne and Pacherie 2005). My view dissolves this debate.

This account of delusions illustrates the central point in the Capacities View of Belief: that evidence-resistance does not entail intractable irrationality. Instead, evidence-resistant beliefs are the result of removable factors interfering with our rational capacities. In this way, the Capacities View of Belief provides a new framework in which to understand real-world evidence-resistance in fine-grained and well-motivated ways.

I call this framework the *Layered Model of Belief Revision.*[4] According to this model,

[4] The Layered Model of Belief Revision could be made precise by assigning weights to the different layers that are at play in belief revision, and by determining what triggers the activation of different constituents of the layers. Ultimately, this could lead to a detailed scientific model of belief revision, one that takes into account Bayesian machinery, motivated reasoning, and external, social factors.

belief revision is regulated by three layers: evidence-responsiveness capacities, internal masks, and external conditions.[5]

This model agrees with traditional epistemology in claiming that beliefs are fundamentally regulated by evidence. This holds in two senses. First, as the Capacities View states, beliefs are constitutively in the province of rational evidence-responsiveness capacities. Second, on this model, factors other than evidence only affect belief revision indirectly, by affecting the operation of evidence-responsiveness capacities or the evidence that serves as input to these capacities.[6]

The second layer consists in other mind-internal factors. These affect belief revision by shaping the evidence that functions as input to evidence-responsiveness capacities. The central family of factors I investigate in the dissertation are motivational factors, which we can construe as desires to hold on to (or avoid) certain beliefs. These desires have a variety of sources, including desires to defend one's self-esteem, desires to pursue difficult courses of action, or desires to maintain a sense of meaningfulness in life, among others.[7] There is work to be done on how attitudes other than desires—such as imaginings and emotions—might mask evidence-responsiveness capacities, functioning as components of this second layer.

The third layer is constituted by mind-external factors.[8] Most obviously, the evidence available in an agent's environment is a component of the third layer. In this way, the social world directly affects the inputs to evidence-responsiveness capacities. In addition, external factors shape elements of the second layer: they shape the emotions, imaginings, desires, and so on that can function as masks on evidence-responsiveness capacities.

[5] Beliefs might also change in ways that have nothing to do with changes in evidence: think, for example, of forgetting. This model does not apply to such cases, which I do not explore in the dissertation.

[6] It is because beliefs are fundamentally regulated by the evidence in these two senses that rational evidence-responsiveness capacities constitute the inner layer in the model.

[7] See Quilty-Dunn 2020 and D. Williams 2021 for discussion.

[8] These constitute the outermost layer because they affect our beliefs through affecting mind-internal factors. In contrast, though motivational factors and the like can affect what evidence agents have in their environment (e.g. by controlling how agents select their environment), they can also affect how we interact with evidence we have without affecting mind-external factors.

Two of the chapters in this dissertation apply this model to specific cases: delusions and resistant social beliefs. In delusions (chapter 3), motivational factors, abnormal experiences, and difficulties suppressing cognitive biases (elements of the second layer) lead to evidence-resistance. In resistant social beliefs (chapter 4), identity-protective desires—resulting from the social networks agents are embedded in—lead to evidence-resistance. In applying the Layered Model in these ways, I illustrate how evidence-resistance is compatible with, and in fact well-explained by, the presence of evidence-responsiveness capacities. Beyond these cases, this model can be put to use to analyze ideological beliefs, religious beliefs, and beliefs in conspiracy theories, among others. I hope to pursue work along these lines in the future.

The Layered Model also has implications for debates in psychology. If we understand epistemic rationality in a Bayesian way, the model allows us to combine the insights of Bayesian theories of belief updating (Tenenbaum et al. 2011) with those of the cognitive dissonance tradition (Mandelbaum 2019). The story Bayesians tell about belief updating captures what happens when the second layer is sufficiently inactive: (bounded) Bayesian updating on the evidence the agent has. Results that Bayesianism struggles to explain—such as the belief polarization effect (Mandelbaum 2019)—are the result of the second layer's operation. On this picture, Bayesian machinery is always present, and plays an important role. All the same, we need resources beyond Bayesianism to describe belief updating.[9]

The Capacities View of Belief and the Layered Model of Belief Revision yield an account of human irrationality that differs in important ways from other recent rationalist proposals. Existing pushback against irrationalist interpretations of findings in psychology has taken a Panglossian perspective: it has involved *denying* appearances of irrationality. Most prominently, Bayesians often ascribe implausible background beliefs (pri-

[9] Note that we can combine a descriptive Bayesian story with one where the ultimate standards of rationality are *not* Bayesian, or where multiple kinds of epistemic rationality figure in our normative theory. My preferred route involves thinking of Bayesianism as describing one kind of epistemic rationality, but not the only one.

ors) to agents, so that rational updating results in seemingly irrational beliefs. Other theorists re-construe the demands of rationality to make room for the claim that apparently irrational responses to evidence are in fact epistemically permissible. This is the strategy that T. Kelly 2008, Dorst 2019, and Begby 2021b take up to argue that biased assimilation, belief polarization, and sticky prejudiced beliefs (respectively) can be rational.

I have no doubt that, as these theorists tell us, we sometimes over-diagnose irrationality. But a version of rationalism that denies all appearances of irrationality is unattractive. It would require us to say that human beings are *actually rational* even when holding on to bizarre beliefs that seem to fly in the face of all available evidence. We would no longer be in a position to dole out epistemic criticism in such cases. In effect, this strategy de-fangs epistemology. It makes epistemic norms powerless at helping us get to a shared view of the world.

In contrast, my version of rationalism leaves room to acknowledge that irrational beliefs are commonplace. Further, it does not require us to abandon stringent standards of rationality.[10] This means that we can continue to criticize bad epistemic behavior, and to promote better epistemic behavior. At the same time, the rationalist claim that rational capacities are present provides a concrete sense in which it is within our ken to arrive at epistemic improvements: doing so is a matter of unmasking rational capacities. A version of rationalism that accepts our fallibility is also one that makes space for aspiration and improvement.

## 1.3 Belief & Sophisticated Agency

The Capacities View of Belief has as a consequence that beliefs are in the space of reasons: because beliefs are evidence-responsive, they are the kinds of attitudes that it is fitting to try to change by offering relevant evidence.[11] Despite placing *beliefs* in the space of

[10] Indeed, the view is for the most part neutral on what these standards ultimately are, allowing for philosophers of different persuasions to slot their preferred views into my account.

[11] In arguing that delusions count as beliefs on such a conception of belief, I aim to show that the space of reasons extends further than often thought, to cover cases of extreme behavioral deviance from the canons

reasons, and thus providing a sense in which rational agency is constitutively involved in belief, this view does not tell us much about to what extent agents self-regulate in response to reasons in sophisticated ways.

Much like the idea that beliefs are constitutively rational, the rationalist idea that many of our beliefs involve sophisticated agency is under threat. The threat emerges from thinking about the role of contextual factors and social structures in how we manage our beliefs. What we believe and how we update our beliefs seem to depend largely on factors that are external to our own epistemic agency, such as what others around us believe, the social costs and benefits of specific beliefs, or whether we are feeling irritable or peppy when discussing a topic. This gives intuitive force to the view that our beliefs are fragile reeds at the mercy of the external world, instead of robust expressions of our epistemic agency.

Two of the chapters in the dissertation (chapters 4 and 5) investigate the role of external factors in belief regulation. They provide resources for defending the idea that sophisticated agency is characteristically involved in human belief.

My view treats sophisticated agency and rationality in asymmetric ways. I defend the claim that rationality is constitutive of belief. But I do not defend the claim that sophisticated agency is constitutive of belief. Instead, I focus on defending the claims that there is room for such agency in belief regulation, and that, if we know where to look, we often find such sophisticated agency at play in adult human belief. This part of the dissertation illustrates the perspectival role of philosophy, i.e., the role of philosophy in helping us see things more clearly, attending, evaluating, and interpreting in better ways (Camp 2019). Specifically, my goal is to develop conceptual tools that help us see the role of individual agency in interactions with evidence that otherwise seem passive.

Chapter 3 considers the idea that not only individual agency, but individual psychology more generally, are irrelevant to explanations of belief maintenance and change. This idea comes out of thinking about the central role of social structures in explaining many

of rationality.

of our socially and politically relevant beliefs. In discussing such beliefs, structuralist theorists have argued that “to focus on individual psychology is to badly misdiagnose how false beliefs persist and spread” O’Connor and Weatherall 2019, p. 7, and that we should not understand the persistence of socially pernicious beliefs “in terms of individual psychological tendencies, such as motivated reasoning” but of “systems and environments” C Thi Nguyen 2021, p. 231. On such views, individual agents’ beliefs are a direct expression of their social environment, leaving little room for agency in belief maintenance.

Against such views, I argue that we need to attend to individual psychology to understand belief maintenance in the social domain. Specifically, I argue that identity-protective reasoning is an important factor in belief maintenance in the social domain. Part of the explanation for resistant social beliefs goes through individuals’ sense of self. Agents resist counter-evidence so as to defend social identities that are central to their sense of self. Against pure structuralism, to understand evidence-resistance, we must appeal to facts about individual agents.

Agents’ evidence-resistant beliefs reveal facts about the social affiliations that matter to them and how they play out in their sense of self, and thereby facts about who they are. In this way, evidence-resistant beliefs reveal the kinds of facts that are plausibly constituents of the agent’s character. Evidence-resistance is not reducible either to external factors or to sub-personal snafus. Instead, it is a consequence of the way in which the cares and desires of social agents manifest themselves in how they interact with evidence.

The claim that belief regulation expresses such aspects of psychology does not exhaust the role for sophisticated agency in belief maintenance. Two other important aspects of agency in belief regulation need accounting for: distinctively *epistemic* agency and agency as self-regulation.

In chapter 5, I aim to provide resources to make room for agency in these two forms. To do so, I draw on work on aesthetic style (Robinson 1985), perspectives (Camp 2019), and modes of agency (C. Thi Nguyen 2020a) to develop a new theoretical tool for describing and explaining our interactions with evidence: *epistemic styles.* An epistemic style is a way of interacting with evidence that expresses an epistemic personality: a unified set of epistemic parameters, such as how much one cares about getting the truth vs. avoiding falsity, which sources one trusts, how much evidence one needs to revise one's beliefs, and so on.

I argue that we can often elegantly explain how people interact with evidence by appeal to the claim that people put on different epistemic styles in response to contextual factors, and that these styles govern their interactions with evidence in relevant contexts.

More strongly, I argue that epistemic styles do a better job than situationist explanations (Fairweather and Alfano 2017) (according to which we can explain interactions with evidence by appeal to situational factors and not individual traits) at capturing interpersonal and contextual variation in how we interact with evidence. For example, a person who adopts a paranoid style—a style for angry minds, expressive of a combination of Cartesian paranoia about what the evidence shows and epistemic risk-seeking (Hofstadter 2012)—and one who takes up a rationalist style—characterized by an adhesion to Bayesian reasoning, conscientious evidence-gathering, and suspicion of testimony (Metz 2021)—are liable to interact with the same evidence very differently, often leading to persistent disagreement and mutual incomprehension. Contextual factors on their own are poorly placed to explain such systematic interpersonal variation.

Explanations of how agents interact with evidence in terms of epistemic style secure a role for distinctively *epistemic* agency. When agents interact with evidence in an epistemic style, such interactions are the result of setting *epistemic* parameters in a specific way, as opposed to the direct effect of non-epistemic factors. Non-epistemic factors enter into the picture by leading agents to set their epistemic parameters in specific ways.[12] In

[12] This raises the question of how the layered model and claims about epistemic styles connect. Epistemic styles characterize how people are actually disposed to interact with evidence. For this reason, they can be understood as describing the surface patterns that result from the factors in the layered model. When the role of motivational factors and emotions (etc.) results in agents' setting their *epistemic* parameters in ways that constitute a unified package, we can appeal to epistemic styles to explain their interactions with evidence.

such cases, agents' interactions with evidence express more than their worldly cares and desires. Most immediately, they express their epistemic agency. They express things like how much they care about getting the truth vs. avoiding falsity, the sources they trust, how they set their evidential thresholds, and whether they prefer simplicity over fit.

Epistemic styles also function as potential sites for agents to self-regulate their interactions with evidence.

At an intellectualist extreme, one can consciously reflect on one's epistemic style, and put the results of such reflection at the service of deliberative control. Specifically, one can engage in systematic reflection about how one ought to set different epistemic parameters, and then devise a plan to get oneself to adopt the corresponding style.

At a less intellectualist level, epistemic styles function as sites for self-regulation when we employ them in actively flexible ways, as opposed to merely shifting fluidly between styles in stimulus-dependent ways (Camp 2022). To a first approximation, we get to flexibility (as opposed to mere fluidity) when we actively adapt our epistemic style to our context in a way that expresses pre-existing aspects of our epistemic personality, as opposed to merely being triggered by that context. In such cases, we make our style genuinely our own. We can, then, be autonomous with respect to our epistemic styles without conscious reflection on them.

In sum, if I am right about the role of epistemic styles in how we interact with evidence, then our interactions with evidence are often expressive of epistemic agency and function as sites for autonomous self-regulation. Epistemic styles make room for robust forms of agency in belief regulation. At the same time, on this view, who we are as (epistemic) agents is deeply shaped by context, and our epistemic agency is compatible with cross-contextual shiftiness. In particular, the involvement of agency does not require long-term character stability of the sort that virtue theorists (Fairweather, L. T. Zagzebski, et al. 2001) often presuppose. In lightly of this, investigations of epistemic agency and epistemic assessment would do well to focus more on styles and not only on virtues.

## 1.4 Conclusion

This dissertation can be read as a defense of the involvement of rational agency in belief in the face of pervasive evidence-resistance and contextual fluidity in our interactions with evidence. More specifically, the dissertation proposes that belief is constitutively evidence-responsive and that, at least in adult humans, belief regulation characteristically involves and expresses epistemic agency, allowing for sophisticated self-regulation. To make the case for this, I draw on empirical evidence and our normative practices to investigate what the concept of belief picks out, and I propose new theoretical notions with which to understand how we manage our beliefs.

This account of belief and belief revision has both theoretical and practical upshots. Theoretically, it is meant to improve our understanding of belief revision and maintenance, i.e., help us construct more accurate models of people's beliefs and interactions with evidence. This, in turn, will help us develop more applicable accounts of epistemic assessment, ones which cover real-world epistemic agents. Beyond that, understanding how rational agency is involved in belief bears on discussions about the place of reasons in a world of causes and on the grounds for the epistemic normativity of belief.

At a practical level, the resources I develop in this dissertation are meant to help us see that there is room for rational engagement where it might (at least, under certain dominant cultural construals (section 1.1)) seem entirely out-of-bounds. This helps us avoid the epistemic injustice of seeing people who we struggle to understand as outside the realm of rational agency. It makes room for new questions about how to rationally engage across deep differences and in the presence of irrational tendencies. And it helps us resist deep pessimism about the viability of democratic institutions in light of evidence-resistance and contextual influences. It is my hope that the combination of theoretical and practical dimensions of this dissertation illustrates how philosophy of mind and social theorizing can productively inform each other.

# CHAPTER 2

# RESISTANT BELIEFS, RESPONSIVE BELIEVERS

**Abstract:** Beliefs can be resistant to evidence. Nonetheless, the orthodox view in epistemology analyzes beliefs as evidence-responsive attitudes. I address this tension by developing an account of belief that does justice to its epistemic role without limiting beliefs to idealized epistemic agents. In doing so, I argue for a capacities-first account of belief: belief requires the capacity for evidence-responsiveness. More precisely, if a subject believes that $p$, then they have the capacity to rationally respond to evidence bearing on $p$. Because capacities for evidence-responsiveness are fallible and may be masked, beliefs can be held in the face of counter-evidence. Indeed, I will argue that our best science of belief supports the claim that evidence-resistant beliefs result from masks on evidence-responsiveness capacities. This account of belief not only allows for resistance to evidence, but provides us with a framework for describing and explaining actual cases of evidence-resistance.

## 2.1 Introduction

Beliefs are often tenaciously held in the face of counter-evidence. For example, we are unlikely to revise political, moral, or religious beliefs (Markus 1986, Leeuwen 2014), beliefs in theories we are committed to (Chinn and Brewer 1993), and beliefs about ourselves and our talents (Pyszczynski, Greenberg, et al. 1985, Gilbert 2006). How do we explain this evidence-resistance? And what does it tell us about the kind of attitude beliefs are?

On the face of it, the evidence-resistance of belief appears to have drastic implications. Specifically, findings from the cognitive science of belief threaten the idea that belief is constitutively evidence-responsive. Given that this idea is orthodox in epistemology, this raises a worrisome tension between the roles which belief plays in epistemology and in cognitive science. Perhaps epistemologists and cognitive scientists are talking about different phenomena when they talk about beliefs. Or perhaps, when they talk of beliefs, epistemologists are talking about attitudes that are rarely realized in human psychology—belief-for-angels, not belief-for-humans.

I argue that this tension is only apparent. The evidence-resistance of belief does not force us to abandon the claim that belief is constitutively evidence-responsive. Indeed, surprisingly, it turns out that findings in the cognitive science of belief *support* that claim—once we properly understand both those findings and what evidence-responsiveness amounts to.

To argue for this view, I will proceed as follows. In section 2.2, I motivate the claim that belief is constitutively evidence-responsive, and argue that other proponents of that claim leave it under-specified. This makes the claim susceptible to two challenges which I outline in subsection 2.2.1. In section 2.3, I put forward my own account of the evidence-responsiveness of belief. My account centers *capacities for evidence-responsiveness*, and incorporates a detailed account of such capacities. I then argue that this account helps us understand what is going on in cases of evidence-resistant belief (section 2.4). For this reason, the account of belief I develop has promising applications in theorizing about phenomena that have long puzzled philosophers of mind, such as delusions, implicit biases, religious beliefs, and ideologies. At the same time, my way of articulating evidence-responsiveness is robust enough to do justice to the epistemic role of belief, fitting with a naturalistic way of spelling out the claim that belief aims at truth (section 2.5). The result is a conception of belief that unifies work in epistemology and cognitive science.

## 2.2 Challenges for Evidence-Responsiveness

The *Evidence-Responsiveness Thesis* holds that what makes a mental state a belief is, in part, its being responsive to evidence.[1] Thus, nothing can be a belief unless it is evidence

[1] Proponents of the evidence-responsiveness thesis include Currie and Ravenscroft 2002, Egan 2008a, Gendler 2008, Helton 2020, Levy 2015, Mandelbaum 2016, Shah 2003, Shah and J. D. Velleman 2005, M.

responsive.

The wide appeal of this claim is best explained by how it accounts for the epistemic role of belief as the central object of epistemic assessment. For instance, D. Velleman 2000 moves from the view that belief aims at truth to claiming that it is a "conceptual truth" that "the belief that $p$ tends to be… reinforced by additional evidence of it, and to be extinguished by evidence against it" (Shah and J. D. Velleman 2005, p. 500).[2] Similarly, Tamar Gendler moves from the claim that apportioning beliefs to the evidence is a normative constraint on belief to excluding attitudes from the belief category because they are not *in fact* evidence-responsive (Gendler 2008, pp. 565–566). Finally, Grace Helton has recently appealed to an epistemic 'ought-implies-can' principle to argue from the epistemic obligation to revise one's beliefs in response to relevant counter-evidence to the view that subjects have the ability to revise their beliefs accordingly (Helton 2020).

Despite the popularity of the evidence-responsiveness thesis, proponents often neglect to specify what evidence-responsiveness amounts to.

Given that beliefs are not always perfectly apportioned to the evidence, statements of the thesis leave room for failures to respond to evidence. Proponents usually grant something to the effect of "[given that] belief can be influenced by evidentially irrelevant processes such as wishful thinking, responsiveness to evidence must be weak enough to leave room for such additional influences" (Shah and J. D. Velleman 2005, p. 500). But they do not usually specify what this "weak enough" evidence-responsiveness amounts to.

They claim merely that beliefs "tend to" (Shah and J. D. Velleman 2005) evidence-responsiveness, are "quickly revisable" (Gendler 2012), or require "the ability" (Helton 2020) to revise.[3] It remains, on such views, unclear what evidence-responsiveness re-

Smith 2003, Leeuwen 2014, and D. Velleman 2000, with notable earlier proponents including D. Davidson 1985, Dennett 1981), and McDowell 1998.

[2] See also Shah 2003 for further discussion, especially fn. 45.

[3] Helton specifies that "the subject is able to revise [beliefs], given her current psychological mechanisms and skills" (Helton 2020, p. 504), and she clarifies this to a larger extent than other proponents of the view (see pp. 513-515 especially). Still, she only specifies that the subject who has it "exemplifies [that ability] on some range of counterfactual circumstances in which her overall psychology is similar to its current state" (Helton 2020, p. 513), without further explaining which counterfactual circumstances are at play.

quires. Does it require rationally responding to evidence most of the time in the actual world? If it does not require rationally responding more of the time, in what circumstances does it require such responses? How much, and what kinds, of evidence-resistance are compatible with evidence-responsiveness? As I will now show, leaving the view underspecified in this way makes it vulnerable to important objections.

### 2.2.1 The extensional and empirical adequacy challenges

An often-noted challenge for the evidence-responsiveness thesis has to do with *extensional adequacy*, i.e., with correctly classifying all and only beliefs as such. The problem is that many of our beliefs are evidence-resistant: we often fail to adjust our beliefs appropriately to relevant evidence. On many senses of "evidence-responsive," evidence-resistant beliefs are not evidence-responsive.Therefore they are counterexamples to the evidence-responsiveness thesis.

This quick argument against the evidence-responsiveness thesis has most explicitly been articulated by Lisa Bortolotti (Bortolotti 2005a, Bortolotti 2005b, Bortolotti 2009). Bortolotti argues that delusions are beliefs (as they are classified in psychiatry (Association 2013)), yet not evidence-responsive. If this is right, then evidence-responsiveness is not necessary for belief, and therefore belief is not constitutively evidence-responsive.

One might resist this argument on the basis of the fact that delusions are strange, unusual attitudes, with their status as beliefs being the object of substantial disagreement (Bortolotti and Miyazono 2015). However, as Bortolotti notes, non-psychiatric evidence-resistant beliefs are ubiquitous. We are unlikely to revise political, moral, or religious beliefs (Markus 1986, Leeuwen 2014), beliefs in theories to which we are committed (Chinn and Brewer 1993), and beliefs about ourselves and our talents (Pyszczynski, Greenberg, et al. 1985, Gilbert 2006). Indeed, Stanley 2015 pursues a line of argument similar to Bortolotti's starting from ideological beliefs instead of delusions.

For this reason, it is not clear how the evidence-responsiveness thesis can correctly classify all ordinary beliefs. The first challenge, then, is to specify what evidence-responsiveness amounts to in a way that correctly classifies ordinary evidence-resistant beliefs.

There is a second, less commonly noticed challenge for the evidence-responsiveness thesis. It corresponds to a second desideratum on a theory of belief, *empirical adequacy*. For a theory of belief to be empirically adequate, it must be compatible with our best scientific findings and generalizations about the cognitive role of belief. This desideratum reflects the fact that beliefs are not only objects of epistemic assessment. They are also objects of study in cognitive science, figuring in well-established empirical generalizations (Porot and Mandelbaum 2020).[4] Troublingly, the evidence-responsiveness thesis is hard to square with a number of key findings in the psychology of belief revision.

Beliefs commonly persist in the face of counter-evidence (belief perseverance (C. Anderson et al. 1980, C. Anderson et al. 1980, Slusher and C. A. Anderson 1989)). Even more worryingly, they sometimes become more entrenched when the subject receives counter-evidence (belief polarization (Festinger et al. 1956, Lord et al. 1979, Liberman and Chaiken 1992, McHoskey 1995, Taber and Lodge 2006)). The evidence-responsiveness thesis needs to accommodate these findings, but it is not clear how it can do so.

Finally, proponents should explain how the view is compatible with robust generalizations about belief. Specifically, they should address the two following generalizations, which, according to Quilty-Dunn and Mandelbaum 2018, are psychological laws of belief: first, “beliefs will generate a negative, motivational, phenomenologically salient discomfort whenever one encounters counterattitudinal evidence,” which we will then be moved to assuage “by any easily available route” (Quilty-Dunn and Mandelbaum 2018, p. 2367); second, if you believe extremely strongly that $p$ and receive information against $p$, then you will (irrationally) increase your belief that $p$ (Festinger et al. 1956).

Easily available routes to alleviate discomfort are likely to include irrationally updat-

[4] Some philosophers reject the view that belief figures in a mature cognitive science (Churchland 1981). Such philosophers will not be worried about empirical adequacy. But they should still worry about extensional adequacy, which involves accommodating intuitive judgments on how to classify attitudes.

ing. Therefore, the first generalization suggests that irrational updating is commonplace. The second generalization suggests that irrational updating on counter-evidence is nearly inevitable for strongly held beliefs.

More generally, if Quilty-Dunn and Mandelbaum 2018 are right, then irrational responses to evidence are not a quirk or occasional blip, but a consequence of the laws that govern belief. It is hard to see how this is compatible with the claim that belief is constitutively evidence-responsive.

The problem of empirical adequacy has special bite for psychofunctionalists, according to whom belief is a functional kind definitionally tied to psychological generalizations about belief (Block and Jerry A. Fodor 1972, Block 1978, Quilty-Dunn and Mandelbaum 2018). If they are right, then no property can be constitutive of belief unless there is a well-established psychological generalization ascribing that property to beliefs. Unless the empirical laws of belief imply that actual beliefs are evidence-responsive, belief cannot be constitutively evidence-responsive.

### 2.3 The Capacities View of Belief

To address the challenges for the evidence-responsiveness thesis above, we need a proper understanding of evidence-responsiveness, one which allows for an extensionally and empirically adequate view of belief—while still satisfying the epistemic motivations for holding that belief is constitutively evidence-responsive. I will now develop a view which meets these challenges:

> **The Capacities View**: Necessarily, if $S$ believes that $p$, then $S$ has the capacity to respond to evidence bearing on $p$ by rationally updating their belief that $p$.[5]

[5] This view assumes, as standard, that belief is an attitude toward a proposition. But it can be adapted to allow belief to have non-propositional contents (Zimmerman 2018) by extending a notion of relevant evidence to non-propositional contents (with accompanying standards for rationally responding to such evidence).

On this view, evidence-responsiveness is a matter of responding to evidence in special conditions—conditions that are suitable for the exercise of our evidence-responsiveness capacities. Attitudes that are not updated in the light of evidence in such conditions do not count as beliefs.

Rationally responding to evidence, as I am using the term, is updating in an epistemically permissible way in response to evidence one has.[6] Possible responses (starting from the belief that $p$) include ceasing to believe that $p$ and coming to believe that not-$p$ or coming to suspend on whether $p$, or increasing or decreasing one's credence or degree of belief in $p$. I remain neutral on both epistemic permissibility and the nature of evidence; you can fill in the details with your preferred accounts.

The Capacities View is one way of spelling out the popular Evidence-Responsiveness Thesis. Its key feature is the appeal to capacities, instead of dispositions (D. Velleman 2000, Gendler 2008) or abilities (Helton 2020), and the fact that I will provide a detailed account of what such capacities involve, unlike other proponents of the thesis have done in the past (section 2.2).

Both appeals to dispositions and to abilities run into problems when it comes to spelling out the evidence-responsiveness thesis. In a natural reading, a disposition to revise requires a good track-record of revision. Much as it would be unnatural to say that someone is disposed to respond stoically to criticism if they start crying almost every time they are criticized, it can sound strange to say that a belief tends to rational revision if it stays put nearly every time counter-evidence is offered. This reading is presupposed by Gendler 2008, who uses the claim that certain attitudes do not change in response to evidence in many conditions in the actual world to exclude those attitudes from the belief cate-

Further, the view is neutral on whether the correct metaphysics of mind is representationalist (Jerry A Fodor 1987) or dispositionalist (Schwitzgebel 2002). If you are a dispositionalist, read this view as stating a counterfactual that subjects must satisfy to count as believing that $p$. If you are a representationsist, read this as constraining the kind of relation to a mental representation that belief involves.

[6] Note the restriction to responding to evidence *one has*. An agent can satisfy the Capacities View while systematically failing to rationally respond to easily available evidence due to failing to gather such evidence. Further, if no evidence bears on $p$—as, on some views, is the case for necessarily true, or necessarily false, propositions—then the conditional trivially holds.

gory. In contrast, articulating the claim in terms of capacities makes clear that evidence-responsiveness does not require beliefs to rationally change in the light of evidence most of the time. Because some beliefs do not rationally change in the light of evidence most of the time, this is a good reason to opt for capacities instead of dispositions.

Abilities are also less well-suited than capacities to articulate the evidence-responsiveness thesis. The exercise of abilities is typically taken to involve trying (Fara 2008). In particular, if an agent responds to evidence when they try, they have the ability to do so. Though my account entails that beliefs require the ability to respond to evidence (in this sense of ability), it also requires subjects to respond to evidence without trying. Given that standard cases of responding to evidence do not involve trying, this provides a reason to opt for capacities instead of abilities.

### 2.3.1 Evidence-responsiveness capacities

I will now explain what evidence-responsiveness capacities amount to. Drawing on Schellenberg 2018's discussion of capacities, I will first make some general claims about capacities, and then apply them to evidence-responsiveness capacities.[7]

Having the capacity to Φ does not imply that one Φs whenever one engages in the relevant activity, or whenever one tries to Φ. For example, having the capacity to run 10k in under 40 minutes does not imply that one always runs at that pace, and having the capacity to score a goal in soccer does not mean that one's every shot at the goal goes in. Indeed, having a capacity does not even require one to succeed reliably, i.e. most of the time that one exercises that capacity. The runner may mostly run at a slower pace, and the soccer player may only occasionally score. Having a capacity is compatible with a track-record of little success.[8]

Having the capacity to Φ involves Φ-ing in specific conditions that suit that capacity.

[7] Though the notion of "capacity" that I will employ is intuitive, "capacity" is a term-of-art, and there are other plausible usages.

[8] Others (e.g. Sosa 2015) understand reliability to allow for a low proportion of success. Nothing hangs on this terminological difference.

For example, what matters to whether one has the capacity to run a 40-minute 10k is whether one would do so when exerting serious effort, well-rested, highly motivated, and so on—even if one would fail when these conditions are not met.[9] Having the capacity to $\Phi$ is a matter of satisfying counterfactuals of the form "If special conditions $C$ were in place, then the subject would successfully $\Phi$."

Applying this discussion to the capacity to rationally respond to evidence $e$, having such a capacity does not require always responding to $e$ when one has $e$. Instead, it involves satisfying the following counterfactual: if one were to receive evidence $e$ in some set of special conditions, one would rationally respond to $e$. More precisely (drawing on Schellenberg 2018's analysis of perceptual capacities):

> **Evidence-Responsiveness Capacities**. A subject $S$ has the capacity to rationally respond to evidence $e$ from overall doxastic state $D$ just in case $S$ would rationally respond to $e$ were the following conditions to hold:
>
> (a) $S$ is in $D$, and has evidence $e$,
>
> (b) $S$ is minimally cognitively capable,
>
> (c) $S$'s brain state shares all relevant aspects with $S$'s brain state in the actual world, i.e. all aspects which that capacity supervenes, or is grounded, on,
>
> (d) no finking, mimicking, or masking occurs.

Condition (a) specifies that, to determine whether a subject has the capacity to respond to evidence $e$ from overall doxastic state $D$, we need only consider instances in which the subject has the relevant evidence and is in the overall doxastic state which the capacity operates on. A subject's overall doxastic state includes beliefs, suspensions, and credences.

[9] At the same time, having the potential to run a 40-minute 10k if one were to train for many months does not suffice for having the capacity. In particular, if one would not do it (under any conditions) as one is right now, one might have the capacity to acquire the capacity to run a 40-minute 10k, but not the capacity to run a 40-minute 10k. See subsection 2.3.2 and Schellenberg 2018 for more discussion that elucidates this distinction.

We can think of the subject's overall doxastic state in a fragmentationist or unificationist framework. On a unificationist reading, we should count *all* of the subject's beliefs (and other attitudes in the doxastic family) as part of the overall doxastic state that these capacities operate on.[10] In contrast, fragmentationists (D. K. Lewis 1982, Egan 2008b, Borgoni et al. 2021) about belief highlight that not all of a subject's beliefs are "always on". We do not bring all the information we have to each and every one of our actions. Instead, different sets of beliefs (different fragments) are activated in different circumstances. On a fragmentationist view of belief, capacities to respond to evidence are capacities to rationally integrate evidence with one's active beliefs. Though I favor fragmentationism, the Capacities View can be spelled out in either way.

Condition (b) says that not responding when asleep, unconscious, or seriously cognitively impaired (e.g. highly sleep-deprived or having a panic attack) is compatible with having the capacity to rationally respond.

Condition (c) tells us that we should not consider counterfactual scenarios in which there has been interference with the capacity's neural basis in determining whether the subject has the capacity here and now. The background assumption here is that evidence-responsiveness capacities, though multiply realizable, are implemented physically. As we come to understand the human brain better, we should come to know what the implementation of these capacities is in us, and therefore be able to identify when there have been the kinds of neural changes that condition (c) excludes.

Condition (d) tells us that how subjects respond where finking, mimicking, or masking are involved is irrelevant to whether they have the capacity. Finking occurs when the conditions for $S$ acquiring or losing a capacity are the very same conditions as that capacity's manifestation conditions. The classic example is Martin 1994's case of a dead wire connected to a device which senses when the wire is about to be touched by a conductor, and makes the wire live in every such circumstance. Mimicking occurs when there is an

[10] I will omit the "other attitudes in the doxastic family" qualification from here on.

interfering factor—something other than the exercise of $S$'s capacity to $\Phi$—in virtue of which $S$ $\Phi$s. The classic case is D. Lewis 1997's example of the Styrofoam glass that the Hater of Styrofoam breaks whenever it is struck, because they detest the sound it makes when struck (see also A. D. Smith 1977, E. W. Prior et al. 1982). Masking occurs when $S$ exercises the capacity to $\Phi$ but does not succeed at $\Phi$-ing, due to the interference of a masking factor. For example, bubble-wrap masks glass's disposition to break when struck (Johnston 1992, Bird 1998).

Applying these points to evidence-responsiveness, condition (d) tells us that, to determine whether a subject has a specific evidence-responsiveness capacity, we can exclude the following: cases in which when they acquire or lose the capacity to rationally respond to evidence at the same time as they receive evidence; cases in which they respond to evidence in virtue of something other than the exercise of their capacities (e.g. a strange neurological fluke); and cases in which a masking factor interferes with the capacity's exercise.

Excluding such cases implies that this is not a *reductive* counterfactual analysis of capacities. I doubt that we can fully specify the conditions under which a subject would successfully exercise a capacity: the literature on dispositions and abilities (Choi and Fara 2018) suggests that counterexamples to such analyses are forever forthcoming. Nevertheless, a counterfactual account of capacities that explicitly rules out such counterexamples remains epistemically useful in determining whether a subject has the target capacity. Specifically, if a subject rationally responds to evidence in many instances where conditions (a)-(c) are met, we have good reason to ascribe the capacity. And, if (a)-(c) are met but the subject does not rationally respond to evidence, then, unless we can detect a specific mask, fink, or mimicking factor, we have grounds to think that the subject lacks the corresponding capacity.

### 2.3.2 Masks on capacities

I turn now to masking cases, which are crucial to my response to the extensional and empirical adequacy challenges. In masking cases, the subject *has* the capacity to respond to evidence, but some factor masks its successful exercise (Johnston 1992, Bird 1998). More specifically, an agent's capacity to rationally respond to evidence $e$ from their overall doxastic state $D$ is masked just in case (i) the agent has the capacity, (ii) conditions (a)-(c) above obtain (i.e., the agent has evidence $e$ and is in overall doxastic state $D$ and not cognitively incapacitated, and there has been no neural tampering) yet (iii) the agent does not rationally respond to $e$.[11] If this factor were removed, the capacity would be successfully exercised, resulting in a rational change of mind.

Legitimate appeals to masks are not *ad hoc*: they are justified by inference to the best explanation. We should appeal to masks when (and only when) the overall behavior we witness is best explained by a masked capacity, and not by the subject lacking the capacity.

Consider the standard example of a mask: bubble wrap around a fragile object. Why should we think that the bubble wrap masks the object's disposition to break when struck, instead of removing it? The answer is that we bubble-wrap glass precisely because it has that disposition (Bird 2007).[12] Taking the glass to remain fragile when it is wrapped explains why we wrap it to *protect* it by *preventing* it from breaking—not to make the glass stronger, so that it ceases to be disposed to break.

Appeals to masks as the best explanation for some behavioral pattern are commonplace. In psychology, appeals to cognitive load to explain why people fail at tasks that require access to working memory (System 2 tasks (Kahneman 2011)), such as probabilistic reasoning, provide good examples. Given subjects' success when they are not under cognitive load, these cases are well-described as ones where subjects have the capacity to engage in probabilistic reasoning, but this capacity is masked by cognitive load. Sim-

[11] This is adapted from Fara 2008's account of masked abilities.

[12] The equivalence between fragility and a disposition to break when struck is a simplification. See (Bird 2007, pp. 18–24) for discussion.

ilarly, in epistemology, appeals to competences are committed to masks. For example, in Sosa 2015's account, factors that put the agent in bad shape function as masks on their competences, which are retained even when they are in bad shape.

As these cases illustrate, despite the traditional denial that there can be intrinsic masks (Choi 2005, D. Cohen and Handfield 2007, Handfield and Bird 2008), factors that are internal to the agent can function as masks. Appeals to intrinsic masks are justified in much the same way as appeals to extrinsic masks: because the best explanation for the presence of some intrinsic feature of an entity is that it prevents the manifestation of one of the entity's dispositions, capacities, or abilities (Ashwell 2010). Where we find a removable feature of a subject that functions in such a way, we are licensed in describing it as a mask.[13]

These points on justified appeals to masks will guide my discussion in the next section, where I will show that real-world evidence-resistant beliefs are the result of masks on evidence-responsiveness capacities.

## 2.4 The Extensional and Empirical Adequacy Challenges Addressed

### 2.4.1 Space for evidence-resistant beliefs

Having the capacity to rationally respond to evidence only requires responding when the conditions discussed in subsection 2.3.1 are met. But these conditions are not always met, making room for failures to rationally respond to evidence.

First, capacities are only activated in specific conditions. A subject may have evidence-responsiveness capacities without being in those conditions. There might be easily available evidence in their environment which they do not have, in which case they are not in the conditions in which the capacity to respond to that evidence would be applied. Or, assuming a fragmentationist model of belief, they may fail to revise a belief despite

[13] Note, however, that a feature that is always present in the actual world might be removable. It is beyond the scope of this paper to detail the modal force of the "can" here; all the masks to which I will appeal are clear cases of masks that can be removed.

having counter-evidence because it is not in the active fragment when they receives that evidence. Or they might be incapacitated when they receive evidence. Such cases are not failures to successfully exercise one's capacity to respond to evidence, but cases where the conditions for responding to relevant evidence are not met.

Second, capacities to rationally respond to evidence need not be reliable in the subject's environment. They may fail most of the time they are exercised. This will be the case if one's environment (including the subject's internal states) is one where it is statistically normal for one or more of the conditions (a)-(d) discussed above to fail to hold. My account therefore leaves space for evidence-resistance while recognizing that it is often the effect of environmental factors, not of the subject's incorrigible irrationality.

### 2.4.2 The extensional adequacy challenge addressed

To explain how evidence-resistant beliefs are compatible with underlying evidence-responsiveness capacities, I will focus on showing that belief perseverance and polarization are compatible with these capacities. I will argue that belief perseverance and polarization are typically explained by motivational factors masking evidence-responsiveness capacities.[14]

The central insight in the literature on cognitive dissonance is that receiving counter-evidence to one's beliefs hurts (i.e., it generates an unpleasant feeling of cognitive dissonance) and it hurts more the more central the threatened beliefs are to the subject (Festinger et al. 1956, Elliot and Devine 1994, Cooper 2007, E. Harmon-Jones and C. Harmon-Jones 2007). Paradigmatic central beliefs for most subjects include those that constitute a positive and stable self-image (beliefs that one is moral, smart, attractive, and so on) (Gilbert 2006), and beliefs that give meaning to one's life, such as those associated with group affiliations or meaningful activities (Pyszczynski, Solomon, et al. 2015). They may also include beliefs that are especially relevant to actions the subject needs to perform (E.

[14] Other masks, such as abnormal perceptual experiences and cognitive biases, also occur. However, following the vast literature on cognitive dissonance, I take motivational factors to be the central culprit for ordinary cases of evidence-resistance, and will therefore focus on establishing that they function as masks.

Harmon-Jones and C. Harmon-Jones 2007).

The role of the unpleasant feeling of cognitive dissonance is motivational: it motivates the subject to find ways to accommodate counter-evidence while maintaining cherished beliefs.[15] The need to reduce negative affect motivates the subject to engage in the psychological work of incorporating evidence so as to maintain those beliefs. The result is belief perseverance and polarization. Subjects do this psychological work in two main ways: through biased assimilation and the belief polarization effect.

In biased assimilation, people receive evidence both for and against their beliefs but *strengthen* their beliefs. In the most-cited study of this phenomenon (Lord et al. 1979), subjects with strong views in favor of, or against, the death penalty were given mixed evidence on its efficacy. Instead of becoming more uncertain of their views, subjects became more strongly convinced of whatever view they initially held.

Counter-evidence to the subject's view on the death penalty threatens the subject's positive self-image, specifically, their beliefs that they are moral and good at reasoning. Subjects are motivated to avoid blows to their self-esteem, and so experience high levels of cognitive dissonance in these cases. To alleviate dissonance without damaging their self-esteem, they are willing to put in significant effort.

This effort takes the form of scrutinizing studies that go against their views in detail, without doing the same with studies that support their views. Because subjects focus on refuting counter-attitudinal studies, they find many arguments against those studies, and very few against the studies supporting their view. This leads them to take their view to be on even more solid ground than they thought before receiving the evidence.

In the belief disconfirmation effect, subjects receive only counter-evidence to their views, and respond to it by becoming more convinced of those views. In the classic discussion of this effect, Festinger et al. 1956 tracked members of a cult that had as its central

[15] Physiological markers of motivation—heightened electrodermal activity, which is associated with activation of the sympathetic nervous system—are present when subjects experience cognitive dissonance (Elkin and Leippe 1986).

tenet that the world would end on December 21, 1954. Members appeared to genuinely believe this: they had quit their jobs, donated their savings to the cult, and started preparing for the end of the world. When the date came and went and the prediction did not come true, many did not abandon the cult. Instead, they became even more strongly attached to the cult and the views it prescribed. They reasoned that aliens had given planet Earth a second chance, and turned to environmentalism to prevent damage to the planet.

As we have seen, subjects are motivated to avoid accepting the conclusion that a view that matters deeply to their identity is wrong. To protect against this, they explain away discomforting evidence—in the case of the belief disconfirmation effect, by changing other views to maintain cherished fixed points. For example, in the cult case above, members came up with an explanation for why the cult's prediction was wrong—namely, that aliens had changed their mind last minute—which allowed them to continue adhering to the cult.[16]

As expected if motivational factors function as masks, if we remove them or alleviate their weight, subjects become more likely to rationally revise in light of the evidence they receive.[17] We can see this by considering the effects of self-affirmation and cognitive load.

Self-affirmations consist in reminders of the subject's important values, skills, or past achievements, i.e. in offering the subject evidence for their positive self-image. Once self-affirmed, it is easier to abandon cherished views while maintaining one's self-worth

[16] There is another mechanism that may underwrite some instances of the belief disconfirmation effect: evidential pre-emption (Begby 2021a). In evidential pre-emption, the subject's belief system leads to the prediction that they will get counter-evidence, at least from certain sources (e.g. sources outside of one's close circle). In such cases, arguably, given background beliefs, properly exercising one's evidence-responsiveness capacities leads to strengthening beliefs when one receives counter-evidence from those sources.

[17] In biased assimilation and the belief polarization effect, subjects may end up rationally responding to *different* evidence. T. Kelly 2008 argues that, given the differential scrutiny to which one subjects the evidence received in cases of biased assimilation, one ends up with a body of evidence that supports becoming more confident in whatever belief one started off with. Once one has such evidence, one exercises one's capacity to rationally respond to it. Similarly, it is widely recognized that how strong counter-evidence is depends on the range of alternative explanations for that evidence: if there are many plausible such explanations, the evidence is weak. For this reason, it may be that, in the belief disconfirmation effect, once subjects have come up with alternative ways of explaining the evidence, they update rationally. In other words, though motivational factors may mask some of subjects' evidence-responsiveness capacities, it may be that they end up exercising different ones.

(Steele 1988). One can admit to having been wrong, or change one's mind in ways that lower self-esteem, while still retaining the beliefs that one is good and competent, because the balance of evidence one has still supports those beliefs. In this way, self-affirmation reduces the subject's motivation to hold on to their views. Cognitive load results from having to maintain multiple items of information in working memory (e.g. when one attempts two different problems at once). Cognitive load impedes effortful cognitive processes that require access to working memory. Interpreting evidence in biased ways, or coming up with alternative explanations for it, is effortful, and therefore hindered by cognitive load (Ditto et al. 1998, Valdesolo and DeSteno 2008).

As predicted by the hypothesis that motivational factors function as masks, we find that biased assimilation and belief disconfirmation (and the resulting belief perseverance or polarization) are attenuated or eliminated by self-affirmation (G. L. Cohen, Aronson, et al. 2000, Reed and Aspinwall 1998, D. A. Sherman et al. 2000, D. K. Sherman and G. L. Cohen 2002) and cognitive load (Ditto et al. 1998, Moreno and Bodenhausen 1999). Specifically, these factors are removable, and, if they are removed, subjects rationally respond to evindece (or get much closer to doing so).[18]

At this point, an objector might point out that the facts I adduced are compatible with the alternative hypothesis that the subject lacks the capacity to respond to evidence, with self-affirmation and increased cognitive load causing the subject to acquire evidence-responsiveness capacities.

In response, I will follow the strategy for justifying masks outlined in in subsection 2.3.2, and argue that the best explanation for the presence and operation of these motivational factors appeals to evidence-responsiveness capacities.

The key question at this juncture is the following: Why would our beliefs be carefully

[18] Removable motivational masks are compatible with the existence of a permanent *psychological immune system* (Gilbert 2006, Mandelbaum 2019, Porot and Mandelbaum 2020, Quilty-Dunn 2020), which functions to defend us against unhappiness and help us maintain stable motivation in the face of a hostile world. Such a system, if it exists, is flexible, and will provide motivation to hold on to different beliefs depending on the context. Specific motivational masks will therefore be removable.

wrapped up in defensive motivational systems if they were not fragile to the evidence? If beliefs were entirely evidence-insensitive, then counter-evidence would be no threat, for it would not lead to one abandoning cherished beliefs. Cognitive dissonance would play no useful function in the cognitive system.

In contrast, suppose beliefs are evidence-responsive in the sense I have detailed. Then, in the absence of motivational factors (and other masks), they will be revised in accordance with the evidence. But such revisions would, in some cases, be unsettling or painful. In those cases, it would be a good idea—from the point of view of maintaining self-esteem and motivation—for motivational factors to kick in and mask those capacities for revision. That is precisely the role of feelings of dissonance: they serve to motivate the subject to seek out alternative ways to accommodate evidence so as to escape discomfort.

Similarly, the best explanation for confirmation bias in evidence-gathering—for why subjects fail to gather available evidence bearing on their beliefs, especially where that evidence might support a discomforting belief or undermine a positive one—appeals to evidence-responsiveness capacities. Here is a good reason to actively avoid gathering evidence: avoiding revising central beliefs which one would revise in the light of that evidence. In contrast, if one lacked the capacity to respond to the evidence, gathering it would make no difference to what one believes. Evidence-avoidance would be puzzling for fully evidence-insensitive attitudes.[19] This supports the claim that the beliefs at stake are evidence-responsive.

Further, the presence of evidence-responsiveness capacities not only explains the existence but also the *modulation* of cognitive dissonance. It explains the fact that, the stronger the counter-evidence to a cherished belief, the more subjects experience cognitive dissonance. Where counter-evidence is weak, it is easy enough for the subject to find ways of

[19] There is an alternative explanation that is compatible with full evidence-insensitivity: perhaps subjects avoid gathering evidence because they do not want to receive evidence of their own evidence-insensitivity. This explanation is inferior to the one I offer. It imputes to subjects beliefs about the degree of evidence-sensitivity of their own beliefs, and it does not match the phenomenology of a range of cases of confirmation bias, which is one of wanting to avoid being forced to accept a view.

integrating it while maintaining their cherished beliefs, and therefore not much motivation is needed to do so. Weak feelings of dissonance suffice. In contrast, when the counter-evidence is strong, it takes significant cognitive maneuvering to avoid revision. Strong feelings of dissonance are needed to provide sufficient motivation. Without such strong feelings motivating difficult psychological work, the subject's evidence-responsiveness capacities would force the abandonment of cherished beliefs. In sum, if subjects have the capacity to rationally respond to evidence bearing on their beliefs, we get an elegant explanation of the role of cognitive dissonance in modulating responses to evidence. More generally, our best understanding of real-world evidence-resistance indicates that such capacities are involved in evidence-resistant beliefs. Real-world evidence-resistance, then, is compatible with the view that belief is constitutively evidence-responsive, understood along the lines of the Capacities View.[20]

This provides new tools for investigating puzzling evidence-resistant attitudes, such as delusions, implicit biases, religious faith, conspiracy theories, and ideological commitments. In providing a concrete test for whether these attitudes are sufficiently evidence-responsive to count as beliefs, my view makes substantive progress over existing views in the literature (section 2.2). The question to ask is whether it is the case that the subject changes their attitude in the specific circumstances articulated in section 2.3.

I conjecture that it is likely that the evidence-resistance of many of these problem cases can be explained in terms of masks—motivational and otherwise—on evidence-responsiveness capacities. Indeed, I have argued that this is true for perhaps the most intractably evidence-resistant of these cases—delusions—in [redacted]. If my argument there succeeds, we have reason to expect the point to also apply to more tractable evidence-resistant beliefs. Such an explanation of their evidence-resistance would advance our understanding of these phenomena, and it would allow us to preserve the intuitive verdict

[20] Cognitive dissonance has also be used to defend the claim that our beliefs are "minimally rational" not in the sense of evidence-responsive but in the sense that "they respond to perceived irrationality by re-establishing coherence" (Ganapini 2020, p. 10). The two views are different in that, on most views of epistemic rationality, re-establishing coherence does not entail rationally responding to evidence.

that these phenomena are genuine beliefs, offering a straightforward justification for the belief-like epistemic assessability of these attitudes.

### 2.4.3 The empirical adequacy challenge addressed

I will now show that my view is compatible with generalizations about belief revision that seem hard to square with the claim that belief is constitutively evidence-responsive.

In subsection 2.2.1, I highlighted two generalizations that seem especially problematic. The first one is: "beliefs will generate a negative, motivational, phenomenologically salient discomfort whenever one encounters counterattitudinal evidence" which we will then be moved to assuage "by any easily available route" (Quilty-Dunn and Mandelbaum 2018, p. 2367). The second one is: if you believe extremely strongly that $p$ and receive information against $p$, then you will increase your belief that $p$ (Festinger et al. 1956).

If the discussion in subsection 2.4.2 is along the right lines, the generalizations that hold of human belief updating are the result of a dual-layered system of belief revision, with evidence-responsiveness capacities as one layer and motivational masks as another. When you believe extremely strongly that $p$, such masks are very likely to be present. When they are present, you will either maintain or increase your belief that $p$, depending on how strongly motivated you are to believe that $p$. For this reason, the second generalization above holds in this model, and it is therefore compatible with the presence of evidence-responsiveness capacities.

In agreement with the first generalization, the motivational layer involved in belief revision is implemented through the generation of discomfort in the face of counter-evidence. And we are indeed moved to assuage this discomfort "by any easily available route." In an optimistic note, my view indicates that the most easily available routes will often be rationally responding to evidence, given the psychological work it takes to respond in other ways.

Equipped with this dual-layer model of belief revision, we can in fact turn concerns about empirical adequacy on their head. Generalizations about belief revision—in particular, about the role and modulation of cognitive dissonance—are not just compatible with the presence of evidence-responsiveness capacities. They are, if the argument in subsection 2.4.2 succeeds, *best explained* by the claim that evidence-responsiveness capacities are present.

Specifically, the generalizations about belief revision that contemporary psychofunctionalists endorse (Quilty-Dunn and Mandelbaum 2018, Porot and Mandelbaum 2020) commit them to evidence-responsiveness capacities underlying belief. Given that psychofunctionalists claim that belief is a functional kind definitionally tied to our best psychological generalizations about belief (Block and Jerry A. Fodor 1972, Block 1978, Quilty-Dunn and Mandelbaum 2018), this implies that they ought to embrace the Capacities View.[21]

This is significant. Traditionally, as described in section 2.2, the motivation for claiming that belief is constitutively evidence-responsive has come from epistemology. Cognitive science is taken to present a serious problem for that view. Against this, I have here shown that taking cognitive science as our primary guide to the nature of belief supports thinking that belief is constitutively evidence-responsive.

## 2.5 Capturing the Epistemic Role of Belief

I will now argue that the Capacities View does justice to the epistemic role of belief. First, it correctly excludes from the belief category attitudes that fall outside the realm of straightforward epistemic assessment. Second, it illuminates a robust sense in which belief can be said to aim at truth, a claim that paves the way for popular explanations of the epistemic assessability of beliefs (D. Velleman 2000).

A desideratum on a notion of belief is that it does not count as beliefs attitudes that do

[21] This is conditional on their acceptance of the generalizations I mentioned. This means that I am not here doing justice to different versions of psychofunctionalism. Nonetheless, the point is significant, given that this is the most developed psychofunctionalist proposal in the contemporary literature.

not play the epistemic role of belief, in particular, attitudes that are not criticizable purely on epistemic grounds (e.g. in virtue of their falsity, or for failing to constitute knowledge).

My view meets this desideratum. Non-doxastic cognitive attitudes—attitudes which present their content as true but which are not beliefs, and which are not striaghtforwardly assessable on purely epistemic grounds—are not evidence-responsive in the capacities sense.[22] To show this, I will focus on the central cases on non-doxastic cognitive attitudes: imaginings and acceptances.

Imaginings are not constitutively evidence-responsive in the capacities sense. Consider imagining that you are on a luxury tropical vacation while receiving decisive counter-evidence: your senses present you with evidence that you are grading in your apartment while it rains outside. In such a case, unless something has gone seriously wrong, your relevant beliefs remain anchored to reality.[23] You believe that you are grading in your apartment and not on a luxury tropical vacation. This reflects the fact that you successfully exercise your capacities to rationally respond to evidence. But the successful exercise of these capacities, made manifest in your beliefs, leaves the imagining intact.

This illustrates that it is possible (in fact, ordinary) to fail to revise one's imaginings in response to counter-evidence in good conditions for the exercise of one's evidence-responsiveness capacities—indeed, in conditions where you successfully exercise those capacities by revising your beliefs. Imaginings, then, are not constitutively evidence-responsive.

Much the same applies to acceptances. Accepting that $p$ is a matter of taking $p$ for

[22] What about non-cognitive attitudes, i.e. attitudes which do not present their content as true (such as desires)? Desires are clearly not evidence-responsive in the same sense as beliefs, i.e., in the sense of constitutively requiring the capacity to revise in light of evidence that what one desires does not hold. Indeed, we typically desire things that do not hold, in full knowledge that they do not hold. A more fitting notion of evidence-responsiveness for desires would appeal to sensitivity to evidence on how likely those desires are to be achieved (as opposed to evidence that their content is true already). I am neutral on whether desires are constitutively evidence-responsive in this sense. See M. Smith 2003 for a view on which they are.

[23] For an account of how beliefs could go that wrong while remaining evidence-responsive, see my [redacted].

granted in a context without believing that $p$.[24] Successfully exercising one's evidence-responsiveness capacities does not (characteristically) touch what one accepts. For example, suppose that you are discussing the future of the European Union with a friend and jointly decide to accept that Catalonia will become independent of Spain, so as to consider what the consequences would be. A third-party who approached with counter-evidence to that claim would be missing the point. Being fully convinced by their counter-evidence—and therefore successfully exercising your evidence-responsiveness capacities to update your beliefs—need not affect your willingness to accept that claim at all.[25]

Indeed, *not* being evidence-responsive may be constitutive of imagining and accepting. In diametric opposition to beliefs, these attitudes play the cognitive role of helping us get away from reality and explore alternative ways things could be. Imagining and accepting are standardly taken to be subject to a decoupling mechanism that insulates them from counter-evidence (Leslie 1987, Perner 1991). If that is right, then there are irremovable barriers to evidence-responsiveness capacities operating on these attitudes, in the sense that, if those barriers were removed, the attitude would cease to be an imagining or acceptance.

Though my view correctly excludes imaginings and desires, there may be constitutively evidence-responsive attitudes other than belief. Suspension (or at least some of its sub-types; see McGrath 2020) is a particularly good candidate for a constitutively evidence-responsive attitude. This is an advantage of the view. To the extent that we think suspension, like belief, is constitutively subject to epistemic standards (Miracchi 2019), it is natural to think that it is also constitutively evidence-responsive. This raises the intriguing possibility of defining a class of doxastic attitudes, including but not lim-

[24] "Acceptance" is a technical term. In one popular usage, "to accept a proposition is to treat it as true" so that acceptance is "a category that includes belief" (Stalnaker 2002, p. 716). I am here using it so that beliefs do *not* count as acceptances. See Van Fraassen 1980 and Bratman 1992 for discussion of acceptances as a distinct attitude type, subject to different norms from the ones that govern belief.

[25] If it does, it will be because you decide you are no longer interested in considering that counterfactual scenario. But we are often interested in considering counterfactual scenarios that evidence suggests to be unlikely.

ited to belief, in terms of the involvement of evidence-responsiveness capacities. And it supports the claim that the necessary condition I place on believing captures the fact that belief is an epistemic attitude.

Another reason to think that the Capacities View captures the epistemic role of belief is that it illuminates the claim that belief aims at truth (or at other epistemic goods, such as knowledge (Williamson 2002)).[26] A popular way of cashing out the "belief aims at truth" metaphor appeals to the the claim that beliefs are regulated by evidence-responsive systems. In David Velleman's words, belief aims at truth in that it is necessarily in the province of systems which "regulate cognitions in ways designed to ensure that they are true, by forming, revising, and extinguishing them in response to evidence and argument" (D. Velleman 2000, p. 253).

My discussion vindicates this view. Specifically, I have offered an account of how, even in cases where how we respond to evidence is decisively influenced by motivational factors, evidence-responsiveness capacities remain involved. Without such an account, the claim that belief is necessarily regulated by evidence-responsive capacities is vulnerable to objections, for there are beliefs for which these capacities are never, or virtually never, successfully employed, making it unclear why we should think they are regulated as Velleman claims (Quilty-Dunn 2020). Such concerns cannot be assuaged simply by noting that regulation by truth-aiming systems "doesn't require belief to be governed by truth-seeking mechanisms alone," (D. Velleman 2000, p. 254). The question is why we should think that truth-aiming systems are involved at all in cases where beliefs are stubbornly evidence-resistant; to address it, one needs an account like the one I provide.

More strongly, my discussion leaves room for arguing that belief itself aims exclusively at truth. Regardless of how human belief is regulated, it is plausible that there could be epistemic agents whose beliefs are fully insulated from motivational factors, and who

[26] I will omit the 'or other epistemic goods' qualification going forward. My discussion applies to any view that (a) takes belief to have an epistemic aim and (b) spells that out in terms of regulation by evidence-responsive systems.

always respond to evidence in epistemically permissible ways. Theists presumably hold that God is such a believer.

If that is right, *pace* psychofunctionalism, the connections we find in humans between motivation and belief change are not constitutive features of belief, but contingent features of human cognitive systems. On such a view, there is an asymmetry between the belief–epistemic aims connection on the one hand, and the belief–motivation link on the other. Belief itself aims exclusively at truth but is, in our case, embedded in cognitive systems that have other aims, such as maintaining motivation.

This is the view I favor. But the Capacities View does not necessitate it. Given the discussion of psychofunctionalism in subsection 2.4.3, the Capacities View is compatible with taking either cognitive science or epistemology as our central guides to the nature of belief.

## 2.6 Conclusion

This paper addressed the question: In the light of pervasive evidence-resistant believing, what is the connection between belief and evidence-responsiveness? I have argued that, counter-intuitively, the cognitive science of belief supports the claim that capacities to respond to evidence are involved in belief. This supports an account of belief—the Capacities View—according to which belief is constitutively underwritten by capacities to respond to evidence. This account provides a clear sense in which belief, unlike imaginings and acceptances, aim at truth. This view provides resources for analysing real-world evidence-resistant beliefs. More importantly, it allows us to acknowledge both the epistemic role of belief and the ways in which beliefs often fail to live up to our epistemic aspirations.

# CHAPTER 3
# DELUSIONAL EVIDENCE-RESPONSIVENESS

**Abstract:** Delusions are deeply evidence-resistant. Patients with delusions are unmoved by evidence that is in direct conflict with the delusion, often responding to such evidence by offering obvious, and strange, confabulations. As a consequence, the standard view is that delusions are not evidence-responsive. This claim has been used as a key argumentative wedge in debates on the nature of delusions. Some have taken delusions to be beliefs and argued that this implies that belief is not constitutively evidence-responsive. Others hold fixed the evidence-responsiveness of belief and take this to show that delusions cannot be beliefs. Against this common assumption, I appeal to a large range of empirical evidence to argue that delusions are evidence-responsive in the sense that subjects have the capacity to respond to evidence on their delusion in rationally permissible ways. The extreme evidence-resistance of delusions is a consequence of powerful masking factors on these capacities, such as strange perceptual experiences, motivational factors, and cognitive biases. This view makes room for holding both that belief is constitutively evidence-responsive and that delusions are beliefs, and it has important implications for the study and treatment of delusions.

## 3.1 Introduction

Delusions are deeply evidence-resistant. Consider, for example, Esmé Weijun Wang's first-personal description of Cotard delusion:

> "In the beginning of my own experience with Cotard's delusion, I woke my husband before sunup… "I'm dead," I said, "and you're dead, and Daphne [the dog] is dead, but now I get to do it over. Don't you see? I have a second

> chance. I can do better now.” C. said, gently, “I think you’re alive.” But this statement, of course, meant nothing. It was his opinion, and I had my solid belief. I can state that the sky is green, but will you see it as such? (Wang 2019, p. 148)[1]

Deep evidence-resistance is close to being definitional of delusions. The Diagnostic and Statistics Manual of the American Psychiatric Association defines delusions as “not amenable to change in light of conflicting evidence” (Association 2013). Patients with delusions are unmoved by evidence that is in direct conflict with the delusion, often responding to such evidence by offering obvious, and strange, confabulations.

As a consequence, the standard view is that delusions are intractable, that is, not evidence-responsive. This claim has been used as a key argumentative wedge in debates on the nature of delusions. Some have taken delusions to be beliefs and argued that this implies that belief is not constitutively evidence-responsive. Others hold fixed the evidence-responsiveness of belief and take this to show that delusions cannot be beliefs (section 3.2).

Against this shared assumption, I appeal to a large range of empirical evidence to argue that delusions are evidence-responsive in the sense that subjects have the capacity to adjust their take on the content of the delusion in rationally permissible ways in response to relevant evidence (section 3.3). This makes room for holding both that belief is constitutively evidence-responsive and that delusions are beliefs, and it has important implications for the study and treatment of delusions and of people with delusions (subsection 3.4.1).

[1] Wang mostly discusses her experiences with the Capgras and Cotard delusion, which are fairly unusual. These delusions arose, in her case, in the context of schizoaffective disorder, which is a diagnosis often accompanied by a wide range of (common and unusual) delusions.

## 3.2 The Intractability Assumption

When a person has a delusion, they seem to lose their grip on reality. They make claims such as "I am dead" (the Cotard delusion (Cotard 1880, A. Young and Leafhead 1996)), "My partner has been replaced by an impostor" (the Capgras delusion (Capgras and Reboul-Lachaux 1994, Pandis et al. 2019)), "Someone else's thoughts are being inserted into my mind" (the thought insertion delusion (Mullins and Spence 2003)). Even when delusions have fairly mundane content—"My partner is cheating on me," (Othello syndrome (Todd and Dewhurst 1955)), "People have it out for me" (persecutory delusion (Freeman 2007)) or "That person is in love with me" (erotomania (H. W. Jordan and Howe 1980))—they share the same feature of appearing disconnected from the available evidence.

More specifically, delusions exhibit extreme evidence-resistance. As the DSM 5 notes, delusions "are not amenable to change in light of conflicting evidence" (Association 2013).[2] Patients maintain their delusions despite being surrounded by strong counter-evidence to them and tend to hold on to their delusions with tenacity, rejecting, dismissing, or explaining away what looks like decisive counter-evidence.

The ways in which people with delusions interact with counter-evidence to their delusions can look deeply puzzling. For example, consider the patient who claimed that his hand belonged to his doctor and who answered, "Ever see a man with three hands?" with: "A hand is the extremity of an arm. Since you have three arms it follows that you must have three hands."(Bisiach 1988, p. 469). Or consider the patient who held on to her delusion that an acquaintance was in love with her even after he told her on the phone that, not only was he not in love with her, but in fact could barely remember who she was (H. W. Jordan and Howe 1980). Patients might receive counter-evidence as one would take up "a bedtime story" (Wang 2019, p. 158). They may claim to be an exception to all

[2] The DSM's definition of delusion has been subject to vigorous criticism, both for failing to distinguish delusions from non-delusional beliefs and for wrongly excluding some delusions from the category (Bortolotti 2018, Max Coltheart 2007). But few contest the claim that deep evidence-resistance is at least present in the vast majority of delusions.

known facts: since one is dead, feeling one's heartbeat and other physical sensations is not evidence that one is alive, in just this one special case (A. Young and Leafhead 1996, p. 158). They may recognize that what they are claiming is "unbelievable," and have the sense that "something is not quite right" when they make delusional claims while holding on to them nonetheless (M.P. Alexander et al. 1979, p. 335).

The evidence-resistance of delusions is so extreme, and so puzzling, that it has led many to doubt whether delusions are beliefs. It is hard to make sense of attitudes that are so deeply evidence-resistant as aiming at the truth, or as hooked to a shared reality that they seek to represent, as beliefs are standardly taken to be. As Andy Egan puts it, "If we think that a certain sort of evidence responsiveness is essential to belief, then, in many cases, we'll be reluctant to say that delusional subjects genuinely believe the contents of their delusions" (Egan 2008a, p. 266). We can summarize this argument as follows:

**The Anti-Doxasticism Argument**

1. The Evidence-Responsiveness View of Belief: Belief is constitutively evidence-responsive.

2. Intractability: Delusions are not evidence-responsive.[3]

3. Anti-Doxasticism: Therefore, delusions are not beliefs.[4]

The Evidence-Responsiveness View of Belief is the orthodox view.[5] It helps make good on the epistemic role of belief as an attitude that aims at truth (B. Williams 1970, D. Velleman 2000) and is constitutively subject to epistemic standards (Burge 2010). If delusions do not meet the evidence-responsiveness benchmark, the anti-doxasticist holds, they cannot be beliefs, and therefore have some other non-doxastic status. Perhaps patients' utterances are empty speech acts (Berrios 1991). Or maybe delusions are attitudes of some different

[3] I owe the term "intractable" to Reimer 2010.

[4] Bortolotti and Miyazono 2015 helpfully center this argument in discussing the philosophical literature on belief.

[5] See Helton 2020 for illuminating discussion.

type: imaginings (Currie and Ravenscroft 2002), acceptances (Frankish 2012, Dub 2017), hybrids between belief and desire or belief and imagination (Egan 2008a), or attitudes towards mental states (Stephens and Graham 2004).

But one person's *modus ponens* is another's *modus tollens.* If one accepts the standard, DSM-endorsed view that delusions are beliefs, their evidence-resistance *prima facie* puts pressure on the idea that beliefs are constitutively evidence-responsive. From this perspective, we get the following argument:

**The Anti-Responsiveness Argument**

1. **Doxasticism**: Delusions are beliefs.
2. **Intractability**: Delusions are not evidence-responsive.
3. **Anti-Responsiveness**: Therefore, belief is not constitutively evidence-responsive.

In a series of influential works, Lisa Bortolotti has forcefully pushed this argument (Bortolotti 2005a, Bortolotti 2005b, Bortolotti 2009). Other defenders of the doxastic conception of delusions (T. J. Bayne and Pacherie 2005, Reimer 2010) also "deny that there is a constitutive connection between belief and evidence" (T. J. Bayne and Pacherie 2005, p. 183).

Doxasticists offer compelling reasons to hold that delusions are beliefs, in addition to the fact that delusions are both intuitively and in clinical practice classified as beliefs.[6]

Delusions appear to be continuous with ordinary believing, which makes it unnatural to think that they are an entirely different kind of attitude. Run-of-the-mill non-delusional beliefs (for example, political and ideological beliefs, and beliefs that play an important role in justifying one's habits or practices) can also be deeply evidence-resistant. And

[6] Note that one can accept doxasticism without holding that a full theory of delusions will be limited to studying delusional beliefs. As phenomenological approaches note, a full understanding of delusions involves understanding the subjective experience of delusions, which in turn may require studying more pervasive changes in the patient's experience of, and perspective on, the world. See Bovet and Parnas 1993, Louis A. Sass 1994, L. Sass et al. 2011, Louis A Sass and Pienkos 2013 for more on such approaches.

there are plenty of borderline cases between stubborn, but non-delusional, beliefs on the one hand and delusions on the other. For example, how should we classify the belief that Hillary Clinton and prominent democrats run a pedophile ring (LaFrance 2020), among other conspiracist beliefs? Such beliefs seem as deeply disconnected from shared reality as many clinical delusions, yet its proponents are not held to be delusional in the clinical sense. The doxastic view elegantly accounts for the continuity between delusions and ordinary beliefs.

Additionally, doxastic views have explanatory power: they account for the fact that we can often interpret subjects' delusion-related behavior by ascribing beliefs. This is most salient in cases where delusions lead to deeply disconcerting behavior, such as stalking someone the patient believes to be in love with them, or attacking a partner who they believe to be an impostor (as 18% of the 260 Capgras patients in Foerstl et al. 1991's review did). But belief ascriptions also have explanatory power when the delusion does not lead to extreme content-congruent behavior: they explain assertions of the content of their delusion, distress and difficulties with coping, and confabulations in response to evidence. The large majority of patients with delusions who are admitted to a psychiatric institution behave in some ways that ascribing a belief in the delusion helps explain (Wessely et al. 1993), supporting the claim that delusions are beliefs.

The debate appears to be at a standstill. At play are two different conceptions of belief and delusions. Doxasticists sever traditionally-held connections between belief and evidence-responsiveness so as to hold the attractive view that delusions are beliefs. Anti-Doxasticists reject that view so as to protect the orthodox connection between belief and evidence-responsiveness. [7]

Despite these deep disagreements, both sides agree on one thing: the intractability

[7] There are parallel debates about other candidate constitutive features of belief, such as inferential integration, action guidance, and reason-giving. On one side, doxasticists use delusions to argue that these are not constitutive features of belief; on the other side, defenders of traditional conceptions of belief appeal to these features to argue that delusions are not beliefs. Discussing these debates is beyond the scope of this paper. See Bortolotti and Miyazono 2015 and Bortolotti 2018 for overviews.

assumption, that is, the claim that delusions are *not* evidence-responsive. And both the Anti-Responsiveness and Anti-Doxasticism arguments rely crucially on this claim. As we have seen, this claim is *prima facie* very plausible: it explains delusion maintenance in the face of counter-evidence and the strangeness of patients' reactions to such evidence. Plausible as it may be, I will argue that we ought to reject it. Though delusions are evidence-resistant, they are not intractable.

### 3.3 Delusions Are Evidence-Responsive

In this section, I will argue against the view that delusions are intractable. Specifically, I will argue that delusions are evidence-responsive in the sense that subjects have the capacity to rationally respond to evidence bearing on the content of their delusion. At the same time, delusions are not *exclusively* regulated in response to evidence; motivational and affective factors (among others) are also involved. This explains why delusions are evidence-resistant, i.e. often maintained in the face of seemingly decisive counter-evidence.[8]

To make the case for this view, I will start by outlining some key facts about mental capacities in general, and capacities to rationally respond to evidence in particular (subsection 3.3.1). This will yield criteria for the possession of such capacities and put me in a position to argue, by appealing to empirical evidence on delusions, that people with delusions have the relevant capacities (subsection 3.3.2). I will then explain why, though people with delusions have such capacities, they so often fail to use them properly, i.e. why they fail to abandon the delusion in response to counter-evidence (subsection 3.3.3).

Before starting, an important clarification is in order. I want to argue that all (or at

[8] Gerrans 2001 argues for the related view that delusions are performance failures of the subject's capacity for pragmatic rationality, i.e. the ability to apply rules of rationality in context. In his view, subjects with delusions have a capacity for pragmatic rationality, and just fail to apply it in a wide range of circumstances. Though this is similar to my view, we are concerned with different capacities: the capacity to respond to evidence, in my case, and the capacity to apply rules of reasoning in context, in his. Further, I am not equating delusions with failures to exercise such capacities, as Gerrans does. Instead, I explain why delusions are evidence-resistant in terms of masking factors on the subject's capacities.

least the vast majority of) delusions are evidence-responsive. But *prima facie* delusions form a highly heterogeneous class. As we have seen, their content can vary widely, from bizarre claims ("I am dead") to completely ordinary ones ("My partner is cheating on me"). Delusions also differ in their effects on action, affect, and the subject's overall vision of the world. While some delusions lead subjects to act out, sometimes in extreme ways (consider the Capgras patient who, taking his father to have been replaced by a robot, decapitated him to find the batteries (Blount 1986, Silva et al. 1989)), many are more behaviorally circumscribed, with the subject continuing their life more-or-less as before despite their delusional claims. At the inferential level, some delusions have a reduced effect on the subject's other beliefs, while others are elaborated into florid delusional theories (M. Davies and Max Coltheart 2000). Finally, delusions are highly heterogeneous in their accompanying psychiatric diagnosis: they appear in cases of schizophrenia (Max Coltheart 2007), localized brain damage (A. W. Young et al. 1992, Ellis and M. B. Lewis 2001), or dementia (Flynn et al. 1991), among other diagnoses, or even no psychiatric diagnosis (Freeman 2006). And this matters to features of the delusion. Delusions in the context of schizophrenia and related disorders tend to be polythematic and elaborated, that is, they tend to cover multiple interrelated themes and have more noticeable effects on inference, affect, and action, whereas delusions arising after localized brain damage tend to be monothematic and circumscribed, that is, focused on a single theme and comparatively insulated from the rest of the patient's cognition, affect, and action (M. Davies and Max Coltheart 2000).

Given such heterogeneity, one may doubt whether delusions are all in the province of the same cognitive mechanisms. If they are not, then one cannot (for example) generalize from studies about delusions in schizophrenia to claims about delusions in the context of localized brain damage, or perhaps even from claims about Capgras delusion to claims about Cotard delusion.

I am not in a position to decisively refute the possibility that different delusions are in the province of radically different cognitive mechanisms, or to consider the whole range of variation within the category of delusions. That said, we should not overstate the threat of heterogeneity. There are compelling models of delusions that propose that they are in the province of the same range of cognitive mechanisms (Appelbaum et al. 1999, Bell et al. 2008) while allowing for variation in the weights assigned to different factors. Indeed, the framework I will articulate for explaining delusion maintenance in the face of counter-evidence is of this sort.

To vindicate the claim that all or the vast majority of delusions are evidence-responsive, I will present data supporting the evidence-responsiveness of delusions with a range of different contents, degrees, and kinds of circumscription, and in the context of different diagnoses. This suffices to make an inductive case for the claim that all (or at least the vast majority of) delusions are evidence-responsive, though, of course, such a case would be bolstered by further research on a wide range of delusions.[9]

### 3.3.1 The capacity to respond to evidence

I will argue that delusions are evidence-responsive in the following sense:

> **Evidence-Responsiveness**: $S$'s attitude towards $p$ is *evidence-responsive* just in case $S$ has the capacity to rationally respond to evidence bearing on $p$.

In this sub-section, I will put forward key facts about the nature of capacities to rationally respond to evidence. This will put me in a position to explore whether people with delusions have such capacities with respect to their delusions.

Rationally responding to evidence consists in changing one's attitude in epistemically permissible ways when one receives evidence: for example, reducing one's degree of belief or abandoning a belief when one receives counter-evidence or increasing one's degree of belief when one receives supporting evidence (to an epistemically permissible extent). I

[9] Thanks to an anonymous referee for pressing me to clarify the scope of my claim.

will here remain neutral on what the epistemically permissible responses to evidence are, a subject of considerable disagreement in epistemology.

In the case of many delusions, there is a broad consensus that many patients' responses are *not* epistemically permissible: patients often fail to respond to counter-evidence to their delusions in epistemically permissible ways. For example, there is wide agreement that one ought to abandon, or at least become less certain in, the claim that one's partner was replaced by an impostor when given evidence that the impostor looks exactly like their partner and can remember many shared life experiences, which Capgras patients often fail to do. It is on the basis of the claim that many of patients' responses to counter-evidence on their delusion are not epistemically permissible that we claim that delusions are evidence-resistant.

I will grant the standard view of what the epistemically permissible responses are in these cases, that is, I will grant that patients often respond in ways that are not epistemically permissible. This is a concessive move to proponents of the intractability view. If one were to claim that patients' interactions with seeming counter-evidence are epistemically permissible, it would follow trivially that delusions are evidence-responsive. In fact, they would count as evidence-responsive in the strong sense that patients *actually appropriately respond* to relevant evidence when they have such evidence.

For the purposes of this paper, I will adopt a factive conception of evidence (Williamson 2002). On this conception, one's non-factive mental states (i.e. mere appearances) do not count as evidence. In other words, I will show that subjects with delusions have the capacity to respond to factive counter-evidence on their delusions. This is also a concessive move to the intractability view. Patients with delusions receive strange perceptual input, which plays a role in causally explaining their delusions (as I will discuss in subsection 3.3.3). It is therefore widely agreed (though rarely emphasized) that patients with delusions have the capacity to respond to non-factive evidence. Showing that they have the capacity to respond to factive evidence bearing on their delusions is more challenging.

It is also more relevant to the debate about whether delusions are beliefs: the capacity to respond to non-factive evidence does not suffice to hook the delusion to shared reality, and therefore is not especially relevant to whether delusions are beliefs.

So far, I have elucidated what rationally responding to evidence bearing on $p$ involves. I will now focus on the most important element of this conception of evidence-responsiveness: the appeal to *capacities*.

I will start with some general facts about what having a capacity involves.[10] Having the capacity to $\Phi$ does not imply that one $\Phi$s whenever one engages in the relevant activity, or whenever one tries to $\Phi$. For example, having the capacity to run 10k in under 40 minutes does not imply that one always runs at that pace, and having the capacity to score a goal does not mean that every shot at the goal goes in. Indeed, as these examples illustrate, having a capacity does not even require one to succeed reliably, i.e. most of the time that one exercises that capacity.

Instead, having the capacity to $\Phi$ involves successfully $\Phi$-ing in specific conditions that suit that capacity. For example, what matters to whether one has the capacity to run a 40-minute 10k is whether one does so when exerting serious effort, not injured, well-rested, highly motivated, and so on—even if one would fail to do so if a single one of these conditions is not met. Having the capacity to $\Phi$ is a matter of satisfying counterfactuals of the form "If special conditions $C$ were in place, then the subject would successfully $\Phi$".

Applying these points to the capacity to rationally respond to evidence $e$, we can conclude that having that capacity does not require always responding to $e$ when one has that evidence. Instead, having the capacity to respond to $e$ involves satisfying the following counterfactual: if one were to receive evidence $e$ in some set of special conditions, one would respond to $e$ in an epistemically permissible way.

What are those special conditions? We are here dealing with a mental capacity: one which operates on mental states, and whose successful exercise consists in yielding a new

[10] My discussion here draws heavily on Schellenberg 2018's discussion of capacities.

set of mental states. As such, one would expect that factors internal to one's cognitive system could interfere with the successful exercise of such a capacity. For this reason, the special conditions in which one will rationally respond to evidence require the right sort of internal environment, one without internal tampering factors. Salient candidates for such tampering include exhaustion, cognitive biases, motivational factors such as the desire to hang on to a belief that is central to one's identity, and affective factors that cloud one's judgment. The special conditions under which someone who has the capacity to rationally respond to evidence $e$ would actually do so are ones where such interfering mental factors are absent. Importantly, these conditions may routinely fail to be met in the actual world. The subject may be too tired or overwhelmed most of the time, or their motivational structure may make them hold extremely tightly to some of their beliefs.

The central upshot is the following: an attitude can both be evidence-responsive in the sense that the subject has the capacity to rationally respond to evidence bearing on its content, and evidence-resistant in the sense that the subject does not change their mind most of the time they receive counter-evidence. That will be the case whenever, for a given individual and attitude, the right conditions for the exercise of the capacity rarely occur in the actual world.

I will argue that this is precisely what we find with delusions: patients with delusions have the capacity to rationally respond to counter-evidence to their delusion but are rarely in the right (internal) conditions to respond to it. To do so, I will, in subsection 3.3.2, appeal to empirical evidence to argue that subjects with delusions would appropriately change their mind in response to evidence bearing on the delusion in a range of conditions, and, in subsection 3.3.3, explain—by appeal to factors that interfere with the exercise of such capacities—why they fail to rationally respond to counter-evidence in many real-world circumstances.

### 3.3.2 In support of the presence of evidence-responsiveness capacities in delusions

In this sub-section, I will give reasons in favor of thinking that subjects have the capacity to rationally respond to evidence bearing on the content of their delusion. My focus will be on showing that that have the capacity to respond to *counter*-evidence. Their capacity to respond to supporting evidence is not commonly questioned.

*Characteristic interactions with counter-evidence*

As discussed in section 3.2, much of the case for the intractability of delusions is based on patients' puzzling responses to evidence. I will now argue that these puzzling responses do not indicate the lack of the capacity to rationally respond to such evidence; in fact, they offer moderate support to the claim that patients have that capacity.

First, note that patients do not *dismiss* the evidence, but *incorporate* it, that is, make efforts to inferentially integrate the evidence with their delusion and background beliefs (Phillip A Garety et al. 2001). This is often occluded in the philosophical literature on the topic, which highlights the fact that patients hold on to their delusion with tenacity while omitting or minimizing the fact that patients make efforts to accommodate the evidence.

Efforts to accommodate evidence can take different forms. Patients may state that they are an exception to generalizations, bite the bullet on implausible conclusions, or contrive stories that explain away the evidence. This case is illustrative:

> We asked her during the period in which she claimed to be dead whether she could feel her heart beat, whether she could feel hot or cold and whether she could feel whether her bladder was full. J.K. said that since she had such feelings even though she was dead they clearly did not represent evidence that she was alive. She said she recognised this was a difficult concept for us to grasp and one which was equally difficult for her to explain, partly because the experience was unique to her and partly because she could not fully un-

> derstand it herself. We then asked J.K. whether she thought we would be able to feel our hearts beat, to feel hunger and so on if we were dead. J.K. said that we wouldn't and that this experience was unique to her. (A. Young and Leafhead 1996, p. 158)[11]

This patient receives what looks like decisive evidence that she is not dead: her heart is beating, she can feel hot or cold, she can feel whether her bladder is full. She does not change her mind in response. But she does engage with the evidence. She recognizes that, in normal cases, this would entail that a person is dead. But, from her point of view, she *knows* that she is dead; so death must look different from usual in her case.

This pattern of reasoning is recognizable. If you observed, or were told about, something that looks like a violation of the laws of nature but were *certain* of what you saw or heard, you might find yourself reasoning in an analogous way. You know what you saw (or that someone else saw it), so the laws of nature must have been violated in that instance. Arguably, belief in miracles is often supported roughly in these ways. More generally, the ways in which patients integrate counter-evidence through confabulations are similar to those of ordinary believers when they receive counter-evidence to cherished beliefs.

Such responses are rationally impermissible (or, at least, I will grant that they are). The fact that patients integrate counter-evidence with their beliefs in rationally impermissible ways does not imply that they have the capacity to integrate such evidence in rationally permissible ways. However, the similarities between these interactions with counter-evidence and others we witness in non-delusional cases are suggestive. To the extent that a capacity to respond to counter-evidence is present in the latter cases, it is plausible to think it is present in cases of delusion too.[12]

[11] This is, of course, just one case. I do not here present data on what fraction of patients with delusions inferentially integrate the evidence in relevantly similar ways. The role of presenting case studies is to make vivid why certain kinds of behavior are indicative of the capacity to respond to counter-evidence. The pervasiveness of those kinds of behavior is a different matter. Thanks to an anonymous referee for pressing me to articulate the role of case studies in the argument.

[12] Thanks to an anonymous referee for pressing me to clarify this point.

Further, patients often understand how the evidence they receive bears on the content of their delusions, that is, what the epistemically permissible responses would be. Here is a striking example:

> E: Isn't that [two families] unusual?
>
> S: It was unbelievable.
>
> E: How do you account for it?
>
> S: I don't know. I've tried to understand it myself and it was virtually impossible.
>
> S: What if I told you I don't believe it?
>
> E: That's perfectly understandable. In fact, when I tell the story, I feel that I'm concocting a story . . . it's not quite right, something is wrong.
>
> E: If someone told you the story what would you think?
>
> S: I would find it extremely hard to believe. I should be defending myself. (M.P. Alexander et al. 1979)

This interaction shows that $S$ understands how the evidence bears on their delusions, i.e. what the rationally permissible responses would be. If, as Ryle noted, "execution and understanding are merely different exercises of knowledge of the tricks of the same trade" (Ryle 1949, p. 55), then the same capacity is employed to judge responses to evidence and to produce such responses. Assuming such a connection between understanding and execution, the fact that the subject appropriately judges responses to evidence indicates that they have the capacity to respond appropriately.

A third characteristic aspect of how subjects with delusions interact with counter-evidence consists in evidence-avoidance, motivated by strong discomfort provoked by counter-evidence.[13] As Esme Weijun Wang writes about her stint with the Cotard delusion:

[13] Though it is hard to get statistics on just how common evidence-avoidance is in delusion patients, em-

> Being dead butted up against the so-called evidence of being alive, and so I grew to avoid that evidence because proof was not a comfort; instead, it pointed to my insanity. (Wang 2019, p. 157)

This behavior looks remarkably similar to the evidence-avoidance we find in run-of-the-mill confirmation bias, i.e. the ordinary tendency to seek out evidence friendly to one's current beliefs, and to avoid evidence against those beliefs (Klayman and Ha 1987, Klayman 1995, and Nickerson 1998). For example, how different is it from avoiding AIDS testing when one fears one might test positive (Dawson et al. 2006, Lerman et al. 2002)? In both cases, the subject avoids gathering evidence where the beliefs they would end up with, if they revised in accordance with that evidence, would be unpleasant ones (namely, beliefs that they have a serious illness).

One upshot is that delusion maintenance in the face of *available* evidence is sometimes a result of the patient failing to gather that evidence, and not of the patient failing to rationally respond to evidence they have. In other words, once we take into account evidence-avoidance, fewer instances of evidence-resistance—i.e. of failing to rationally respond to evidence one has—are left to explain.

Further, this kind of evidence-avoidant behavior is a sign of the capacity to respond to evidence. The reason is the following. Wanting to avoid revising beliefs one would end up revising if one acquired certain evidence is a good reason to avoid gathering it. This is sometimes explicitly recognized from a first-person perspective. For example, in his account of living with (and overcoming) grandiose and paranoid delusions, Robert Chapman writes that he "was afraid to check with reality for fear that [his] ideas might be deflated and [his] sense of having a useful and meaningful direction in pursuing these might be demolished." (Chapman 2002, p. 547). In contrast, if one were not even able to respond to the evidence, gathering it would make no difference to what one believes.

pirically well-supported models (e.g.Freeman, P. Garety, et al. 2001's model, which focuses on persecutory delusions) ascribe evidence-avoidance a significant role in delusion maintenance. This suggests that it is a pervasive feature of delusions.

Evidence-avoidance would be puzzling for intractable attitudes. [14]

*The bias against disconfirming evidence (BADE)*

A problem for the view that delusions are evidence-responsive is that people with delusions display a bias against counter-evidence (once they have gathered it). In this subsection, I will consider studies on this phenomenon and argue that this bias against counter-evidence is compatible with the capacity to rationally respond to counter-evidence.

In a series of studies, Todd Woodward and collaborators have shown that people with schizophrenia (Moritz and Woodward 2006, Woodward, Moritz, Menon, et al. 2008) and with schizotypal traits (Woodward, Buchy, et al. 2007), who are more likely to experience delusions, display a *bias against disconfirming evidence* (BADE): they adjust their beliefs to counter-evidence much less than other subjects.[15] This bias is restricted to strongly held beliefs (Woodward, Moritz, Menon, et al. 2008), but domain-general, that is, not restricted to delusional topics. The authors of these studies hypothesize that a BADE is a contributing factor to the maintenance of delusions (Woodward, Moritz, Cuttler, et al. 2006).

These findings may seem hard to square with the capacity to rationally respond to such evidence. Indeed, the authors of these studies describe people with a bias against disconfirming evidence as "less *able* to revise false convictions in general" (Moritz and Woodward 2006, p. 158) and "generally *impaired* in their ability to integrate disconfirma-

[14] Note two complications. First, there is an alternative explanation that is compatible with full evidence-insensitivity: perhaps subjects avoid gathering evidence because they don't want to receive evidence of their own evidence-insensitivity (thanks to Andy Egan for suggesting this alternative explanation). But this explanation is inferior to the one I propose. It imputes to subjects beliefs about the degree of evidence-sensitivity of their own beliefs. And it does not match the phenomenology of confirmation bias, which is one of wanting to avoid being forced into a view one dislikes, or first-personal descriptions like the ones above. Second, this does not establish that patients have the capacity to respond to evidence to a rationally permissible extent; perhaps they have the capacity to respond by adjusting to *some* extent in the right direction, but not to a rationally permissible extent (thanks to an anonymous referee for pointing this out). However, it does indicate that counter-evidence can have some substantial effects on patients, which sits poorly with the intractability assumption. I will shortly discuss some reasons why patients with delusions may fail to respond to evidence to a sufficient extent, while still adjusting in the right direction.

[15] Thanks for an anonymous referee for bringing these studies to my attention.

tory evidence" (Woodward, Moritz, Menon, et al. 2008, p. 268) (my italics).

In my view, these are inaccurate descriptions of the implications of these findings. A bias against disconfirming evidence does not imply lacking the capacity to respond to such evidence, as I will now argue.

First, given the domain-generality of the bias, and the fact that it is found in the non-clinical population, if a bias against disconfirming evidence implies lacking the capacity to respond to such evidence, lacking capacities to respond to counter-evidence would be widespread. People with schizophrenia and schizotypal traits, most of whom do not have delusions, would lack the capacity to respond to counter-evidence to all of their strongly held beliefs. This is a very strong, and *prima facie* implausible, claim.

Further,other facts about how these subjects interact with evidence are hard to square with lacking the capacity to respond to counter-evidence. They do as well as other subjects when it comes to responding to supporting evidence. They adjust their beliefs to some extent—just an insufficient one—in light of disconfirming evidence in the studies. And, in other contexts, they show a greater tendency to over-adjust to disconfirmatory evidence (Moritz and Woodward 2005), so that a bias against disconfirming evidence is only manifest in some contexts. These facts sit awkwardly with the idea that they lack the capacity to rationally respond to counter-evidence. Instead, they suggest that these subjects have the capacity to respond to (both supporting and disconfirming) evidence on their beliefs, where that capacity is masked in *some* instances in which they receive counter-evidence to strongly held beliefs.

In fact, other work by the same team suggests a plausible explanation for the bias against disconfirming evidence that does not impute lacking the capacity to respond to counter-evidence. All of us—subject to this bias or not—are motivated to find ways of accommodating counter-evidence to our strongly held beliefs without abandoning them (Gilbert 2006, Cooper 2007, E. Harmon-Jones and C. Harmon-Jones 2007). Now, one would expect interpersonal and contextual variation in how many ways of accommodating counter-evidence one comes up with, i.e. in how many alternative explanations for that evidence (other than the falsity of one's cherished beliefs) one constructs. The more alternative ways of accommodating counter-evidence one can come up with, the less that evidence will seem to discredit one's strongly held beliefs.

We have reason to think that the very same subjects who display a bias against disconfirming evidence come up with more ways of accommodating counter-evidence. This is because they display a *liberal acceptance bias*: they tend to give high plausibility ratings to a wide range of alternative views, including views that common sense would immediately dismiss (Moritz and Woodward 2004, Moritz and Woodward 2005).[16] [17]

Let's put these two points together. Like all of us, people who display a bias against disconfirming evidence are motivated to generate alternative explanations for that evidence. Due to their liberal acceptance bias, they will accept more of those as plausible than other members of the population. This leads them to adjust their beliefs to counter-evidence less, i.e. to display a bias against disconfirming evidence. From their perspective, the counter-evidence does not look decisive because it can be explained away in a wide range of ways.

If this is right, then we have a good explanation for the bias against disconfirming evidence that is compatible with people with this bias having the capacity to respond to counter-evidence to strongly held beliefs in rationally permissible ways. People with a BADE have the capacity to rationally respond to counter-evidence, but it is masked by (1) their motivation to hold on to those beliefs, which leads to generating more alternative explanations for the counter-evidence, and (2) a liberal acceptance bias, which makes them

[16] One might worry that this just pushes intractability one step back, to the intractability of these alternative views: why do they accept views that common sense would immediately dismiss? But liberal acceptance does not require failing to respond to counter-evidence to these implausible alternatives. More plausibly, patients simply fail to *gather* such evidence, in that they may fail to retrieve it from memory and are likely not to receive such counter-evidence from their environment at the moment of acceptance.

[17] These findings cohere with phenomenological descriptions of delusion in schizophrenia, which highlight disturbances in background or bedrock certainties Rhodes and Gipps 2008, "a change in the totality of understandable connections"(Jaspers 1963, p. 97), or "a mutation of the ontological framework of experience itself"(Louis A Sass and Pienkos 2013, p. 633) as factors in the formation and maintenance of delusions. One would expect such disturbances to be reflected in accepting options that common sense dismisses.

fail to rule out some of those explanations which common sense would exclude.

More strongly, one may think that, taking into account the wider range of alternative explanations they consider, people with a BADE respond rationally to counter-evidence. Specifically, if one thinks that how confident one should be in a belief in the face of relevant evidence depends on the space of alternative explanations for that evidence of which one is aware, then displaying a BADE may be rational given the wider range of alternative explanations considered.

### *Delusion remission and cognitive behavioral therapy*

So far, we have been looking at how the behavior of people with delusions when they do *not* change their mind in *prima facie* epistemically permissible ways in response to counter-evidence contains markers of the capacity to respond to evidence. But patients sometimes abandon their delusions in response to counter-evidence or come to hold them less strongly. In this section, I will discuss such instances and consider their significance for whether delusions are evidence-responsive.

First, in some cases, the strength with which a delusion is held wanes in response to counter-evidence, with the subject eventually abandoning the delusion:

> LU was asked whether she had ever seen a dead person before, and if so how she had known that the person was dead. LU responded that after her grandmother's death she had viewed her grandmother, and that she knew her grandmother was dead because her eyes were closed and she was motionless. LU acknowledged that the fact that she herself was moving and talking was inconsistent with the typical characteristics of dead people, and she subsequently expressed some uncertainty about her beliefs. Within a week of the initial neuropsychological assessment, her delusion appeared to have completely resolved. (R. P. McKay and Cipolotti 2007, p. 353)

Giving LU evidence against the claim that she is dead leads her to, first, reduce the degree to which she felt certain of that claim, and then to abandon it altogether. This delusion is evidence-responsive: LU *actually responds* to counter-evidence to the delusion, and therefore trivially has the capacity to do so.

The fact that patients' confidence in their delusions varies also indicates that patients have and exercise capacities to rationally respond to counter-evidence. Conviction can vary even over the course of one day, with a predictor for reduced confidence being interactions with other people (Myin-Germeys et al. 2001). This may be because others provide counter-evidence to the delusion, suggesting that people with delusions are at least sometimes swayed by counter-evidence.[18][19]

Further, delusions are typically not chronic conditions: they tend to be eventually abandoned, at least when they do not arise in the context of neurodegenerative disease. Once the delusion is in remission, subjects exercise (and therefore have) the capacity to rationally respond to evidence. It is unlikely that they lose this capacity during the delusional period and then re-acquire it. In such cases, "it is reasonable to assume that the neuroanatomical basis of... competence is unimpaired" (Gerrans 2001, p. 166).

Notably, first-personal descriptions of coming to abandon a delusion often emphasize the effects of considering counter-evidence. Consider these two cases, the first one of Capgras delusion remission, and the second one of remission of a range of persecutory delusions:

> I've started going through it, and seeing what could possibly happen and what couldn't happen. That was wrong, that couldn't happen. Even though it has happened it couldn't. Mary couldn't suddenly disappear from the room, so there must be an explanation for it. The lady knows me way back. She could say things that happened 40 years go, and I wonder where she gets them from.

[18] This study considers only patients with schizophrenia.

[19] As seen in our discussion of the bias against disconfirming evidence, they are typically swayed to a lesser extent than people without schizophrenia. But, as discussed there, this is no objection to the claim that they have and exercise the capacity to rationally respond to evidence in such cases.

> … And then I worked it out and I've wondered if it's Mary all the time. It's nobody else. (M. Turner and M. Coltheart 2010, p. 371)

> When a delusion is stacked against a conscious awareness of reality and rationality, the delusion falls apart. While the delusive ideas disintegrate, their pretense is revealed. I put on my detective cap. I would test out arguments. I tried to develop the strongest arguments possible against the falsehoods. I made a list of all the rational alternatives that I could think of. I looked for evidence for what really was happening and what really wasn't happening. I asked myself, "How do I know this?" Did I actually see it or just a "sign" of it? Did I really hear it, or could I have misinterpreted what I heard? Did I smell, taste, or feel it? Did someone tell me this? Is most of my evidence beyond my senses or interpretations of signs and symbols? I tried to test reality in terms of the here and now transactions with other people rather than assuming what their supposed intentions were or predicting what would happen.(Chapman 2002, p. 551)

In cases of remission like these, patients respond to evidence (from their senses, testimony, and memory). Considering counter-evidence leads them to abandon their delusion. They have the capacity to respond to evidence, and successfully exercise it to abandon the delusion.

These first-personal descriptions are indicative of the fact that offering counter-evidence has the potential to make the patient abandon the delusion. This lies behind the effectiveness of cognitive-behavioral therapy (CBT), which provides a particularly compelling case for the view that delusions are evidence-responsive. Cognitive-behavioral therapy directly guides the patient toward collecting and assessing evidence bearing on their delusion. It centers on changing patients' beliefs about events they have experienced so as to change their responses to those events. More specifically, it focuses on *cognitive restructuring*, i.e., on examining and challenging maladaptive thought patterns and helping the patient establish more adaptive ones (Dozois and Dobson 2010, p. 11) and on *reality testing*, that is, assessing whether beliefs fit the evidence and replacing distorted beliefs with more realistic ones (Kendall and Bemis 1983). The second quote on delusion remission above is illustrative of the process (though, remarkably, the patient himself designed and conducted the process).

Though the philosophical significance of cognitive-behavioral therapy has been insufficiently explored, doxasticists (Bortolotti 2009 and T. J. Bayne and Pacherie 2005) have appealed to it in arguing for the claim that delusions are beliefs. They argue that our theories of delusion ought to be consistent with best clinical practice, in which "the therapist treats the delusional patient as a believer of $p$, and he or she gently invites the patient to question whether $p$ is the thing that ought to be believed" (T. J. Bayne and Pacherie 2005, p. 185). Consistency with such practice, they argue, requires classifying delusions as beliefs, as opposed to claiming that the therapist is mistaken in their ascription.

Note that such a practice also involves the therapist assuming that the patient has the capacity to respond to counter-evidence: the therapist assumes that the patient can "question whether $p$ is the thing that ought to be believed"(T. J. Bayne and Pacherie 2005, p. 185) and abandon their belief that $p$ if it is not. By the same token, then, doxasticists ought to accept the evidence-responsiveness of delusions. This is significant because, as we saw in section 3.2, they endorse the intractability assumption.

The effectiveness of cognitive-behavioral therapy can be marshaled in an additional argument for delusional evidence-responsiveness. Specifically, the best explanation for successful instances of cognitive behavioral therapy involves the target attitudes being evidence-responsive. Successful CBT requires the patient to correctly assess the significance of the evidence they receive, and adjust their delusion and related beliefs accordingly. Properly guided by the therapist, the patient *actually responds* to counter-evidence, which establishes that they have the capacity to do so at that moment. The best explanation for this fact is that they had that capacity before entering the therapeutic context, but were failing to exercise it properly.[20]

CBT is effective as a treatment for delusions. Recent meta-analyses concur that "targeted individualized CBT for delusions and hallucinations is effective" (T. Lincoln and Peters 2018; see also Gaag et al. 2014). These studies cover both delusions in the context of schizophrenia and other clinical contexts.[21] They establish that therapeutic approaches with a "focus on cognitive reframing and reality testing" (T. Lincoln and Peters 2018, p. 67)—i.e., on getting patients to assess the content of the delusion based on evidence—are effective, exhibiting significant benefits on "subjective strength of conviction, reactions to and acting on beliefs" (T. Lincoln and Peters 2018, p. 67). In other words, patients have been found to respond to evidence against their delusions by abandoning or reducing their credence in their delusions.

We have seen that, for cognitive-behavioral therapy to work on an attitude, that attitude must be evidence-responsive, and that CBT works on many delusions. The upshot is that many delusions are evidence-responsive.[22]

### 3.3.3 Explaining delusional evidence-resistance

In the last sub-section, I presented a wide range of reasons to think that delusions are evidence-responsive. Still, patients hold on to their delusions with a tenacity that seems hard to understand. They do not employ their capacities to respond to evidence often or easily. This requires explanation: what is keeping them from successfully exercising these capacities? Three families of factors are promising candidates for explaining evidence-

[20] The alternative is that they acquire this capacity in therapy, but lacked it beforehand. This alternative implies deep changes to the patient's cognitive architecture over the course of a small number of therapy sessions, which is implausible.

[21] For reviews that focus specifically on schizophrenia, and establish that CBT is effective in treating delusions in schizophrenia, see Sarin et al. 2011 Wykes et al. 2008 Turkington et al. 2006.

[22] This result should not be over-stated. CBT is *relatively* effective as a treatment for delusions, but this is in part because delusions are still poorly understood, and, as a result, we lack highly effective treatments. Indeed, most of the effect sizes found are in the small-to-medium range (Gaag et al. 2014), and there are many cases of delusions where this treatment does not work at all. That said, when it works, it is in part because the delusion is evidence-responsive. When it does not, this is likely due to the kinds of factors I will outline in subsection 3.3.3.

resistance in delusions. These are continued strange experiences reinforcing the delusion, motivational factors, and automatic reasoning biases. In this section, I will show that these factors and their interaction with evidence-responsiveness capacities can explain a wide range of evidence-resistant behavior in patients with delusions.

Take strange experiences. Many delusions are formed in response to strange experiences. For example, the best explanation for Capgras delusion ("My partner has been replaced by an impostor") formation appeals to a deficit in the visual processing of faces that is the mirror image of prosopagnosia.[23] In prosopagnosia, people lose the ability to recognize faces but still experience the affective responses they would usually have in response to the faces of loved ones, as measured by, for example, skin conductance tests (Bruce and Andy Young 1986). Conversely, Capgras patients' facial recognition systems are intact, but they lack the usual affective responses they would have upon seeing loved ones (Ellis and A. W. Young 1990, Ellis and M. B. Lewis 2001). This generates the sense that something is deeply wrong. It appears to the patient that something has changed in the person, and this needs accounting for. The view that they have been replaced by an impostor is a response to this abnormal experience.[24]

Though the strange experience does not suffice to explain delusion formation, it is widely agreed that it is a crucial causal factor.[25] Further, this is not a one-off experience.

[23] This explanation paradigmatically applies to cases of Capgras that occur without a schizophrenia diagnosis, but Max Coltheart 2007 argue that it also extends to such delusions in the context of schizophrenia. Similar explanations in terms of localized brain damage causing perceptual distortions have also been explored for other monothematic delusions.

[24] This brief explanation omits many points of disagreement about Capgras formation. For example, some accounts hold that the first step in delusion formation is not a conscious experience but the mere lack of an autonomic response (Max Coltheart 2005). And some hold that the patient *endorses* the experience of seeing their partner as an impostor (T. Bayne and Pacherie 2004), instead of adopting it as an *explanation* for the experience that something is off in the interaction (Maher 1999, Stone and A. W. Young 1997). Others still think that the perceptual effects we find in Capgras are the result of a top-down disturbance (Campbell 2001: the experience is the result of the delusion, and not the other way around. Which of these views is correct is an interesting question, but it does not affect the point that delusion formation involves a strange experience, which is the key point for my purposes.

[25] Most theorists agree that there are other factors involved in the formation of the delusion, such as a deficit in hypothesis evaluation (Max Coltheart 2007), reasoning biases such as the tendency to jump to conclusions (Philippa A. Garety and Freeman 1999), a liberal acceptance bias (Moritz and Woodward 2004, Moritz and Woodward 2005), or a bias toward privilege explanatory adequacy (i.e. privileging how well the experience is explained by the hypothesis over how probable the hypothesis antecedently is (R. McKay

It is sustained by cortical damage that does not just go away. As a result, the patient's unsettling sense of unfamiliarity when looking at the face of a loved one is continually reinforced: whenever the Capgras patient looks at their loved one's face, they experience the same disturbing lack of affect. They *constantly* receive apparent evidence that supports their delusion. This comes into conflict with the actual evidence they get from talking with other people or recalling facts they know. Responding to that evidence in rationally permissible ways is difficult in the face of such strange private experiences. In other words, these strange experiences explain why exercising the capacity to respond to counter-evidence does not result in the abandonment of the delusion.[26]

To get a clear sense of the difficulty in abandoning a delusion in such circumstances, consider the following case of remission from anosognosia (the delusional failure to acknowledge illness or impairment (A. M. A. Davies and M. Davies 2009)):

> E: What was the consequence of the stroke?
>
> HS: The left hand here is dead and the left leg was pretty much.
>
> HS: (later): I still feel as if when I am in a room and I have to get up and go walking… I just feel like I should be able to.
>
> E: You have a belief that you could actually do that?
>
> HS: I do not have a belief, just the exact opposite. I just have the feeling that sometimes I feel like I can get up and do something and I have to tell myself 'no I can't'. (Chatterjee 1996, p. 227)

The patient believes that his left leg is paralyzed: he knows he had a stroke that left him paralyzed on the left side of the body. That said, he still experiences an inclination towards the delusional claim that he can move around as before the stroke. This inclination seems to partly result from strange, illusory perceptual experiences of their paralyzed

2012)).

[26] See Corlett et al. 2009 for an account of delusion maintenance that emphasizes the role of such reinforcement.

limb moving as intended, where these experiences result from the impairment of sensation, attention, and motor control (A. M. A. Davies and M. Davies 2009). It takes continued suppression and monitoring efforts for the patient not to be taken in by this powerful feeling. Given this fact, it is not surprising that patients sometimes temporarily become less certain of their delusion when given evidence against it, but then return to strongly believing it.

Following the philosophical literature on delusions, I have so far focused on the role of strange experiences in monothematic delusions that result from localized brain damage. But the same point applies to delusions characteristic of schizophrenia. They also appear to be partly formed and sustained by abnormal perceptual experiences. Due to dopamine deregulation, patients with schizophrenia attribute salience inappropriately to internal and external stimuli (Kapur 2003). This results in a strange experience of the world writ large (as the phenomenological tradition has long recognized), one characterized by a sense of tension and hidden meaning (Fuchs 2005, Louis A Sass and Pienkos 2013). Again, this sense of strangeness and significance is continually reinforced in each and every perceptual experience, coming into conflict with evidence received via testimony or memory.

Strange perceptual experiences, then, are an important factor in explaining why people with delusions maintain them in the face of counter-evidence, despite having the capacity to respond to such evidence. That said, not all people with delusions have such strange experiences (Bell et al. 2008). Such experiences cannot be the only factor masking evidence-responsiveness capacities.

Motivational factors (such as the desire to hang on to a pleasant delusion or a compelling explanation for one's strange experiences) are another such factor. The role of motivational factors in delusion has long been recognized. Indeed, traditional psychodynamic theories of delusion formation held that these were formed exclusively due to motivational factors. Such theories claim, for example, that delusions of persecution are part of a defense mechanism to avoid attributing negative events to oneself, and that

Capgras delusion arises to make sexual desire for an inappropriate target (e.g. a parent) acceptable.

These views do not hold up to scrutiny, given evidence of brain damage, anomalous experiences, and reasoning biases and deficits in delusions. Nonetheless, it remains very plausible that motivational factors play some role in the maintenance of delusions, helping explain why subjects' evidence-responsiveness capacities do not lead to revision in the face of counter-evidence.

This is especially plausible when it comes to delusions with a content that is self-aggrandizing—such as delusions of grandeur, or erotomania—or simply more positive than evidence warrants—such as anosognosia or the reverse Othello syndrome (the delusion that one's partner continues to be faithful, despite overwhelming evidence to the contrary). Sophisticated motivational accounts of such delusions have been developed (e.g. Ramachandran 1996's account of anosognosia). Similarly, persecutory delusions seem to be maintained in part because attributing negative events to external causes allows the patient to avoid negative beliefs about the self (Bentall et al. 2001).

The basic idea here—explored in detail in accounts that see self-deception and delusion as overlapping—is that holding on to such views supports one's self-esteem or self-image, motivating the subject to maintain them.[27] This inclines the subject to process evidence in biased ways, confabulating (Turnbull et al. 2004) and more generally finding elaborate ways of fitting counter-evidence by adjusting some of their other beliefs.

The role of motivational factors is not restricted to positive delusions. As cognitive dissonance theory (Festinger et al. 1956, Elliot and Devine 1994, Cooper 2007, E. Harmon-Jones and C. Harmon-Jones 2007) tells us, receiving counter-evidence to our beliefs (positive or not) has negative valence. We are motivated to alleviate cognitive dissonance by re-establishing coherence among our beliefs, now including the counter-evidence. How we do so depends on the centrality of the beliefs under attack: the more central or strongly

[27] See the essays in T. Bayne and Fernández 2010 for discussion of the relationship between delusion and self-deception.

held they are, the more likely we are to alleviate dissonance while maintaining those beliefs.

Two facts about delusions, regardless of whether their content is positive, should make us expect patients to be highly motivated to maintain them. First, delusions provide an explanation for an unusual experience, without which the patient would feel at sea in the world.[28] Second, in the case of delusions, admitting that one is wrong is admitting that something has gone seriously amiss with one's cognition, a conclusion that would likely ravage one's self-esteem. For these reasons, as seen in the discussion of the bias against disconfirming evidence above, we should expect patients to expend significant effort at generating alternative explanations for counter-evidence to their delusions. In this way, motivational factors would contribute to delusion maintenance by leading patients to come up with a wide range of ways of accommodating counter-evidence.

Finally, successfully exercising one's capacity to respond to evidence can be effortful, requiring one to overcome or suppress automatic reasoning biases. Where such biases are operative, one will fail to rationally respond to evidence, despite having the capacity to do so. Reasoning biases function as masks on the subject's cognitive capacities: they are "thinking distortions and processing preferences rather than performance deficits and limitations of mental capacity" (Moritz and Woodward 2007, p. 619). In other words: exhibiting reasoning biases does not imply that one lacks the capacity to reason rationally, only that exercising that capacity will require effort to overcome or suppress those biases.

As seen, people with delusions exhibit several reasoning biases (including attribution biases, the tendency to jump to conclusions, and a bias towards observational adequacy) to a greater extent than control subjects (Broome et al. 2007, Philippa A. Garety and Freeman 1999, R. McKay 2012). To respond to evidence, they will need to suppress these biases, which is cognitively effortful and requires motivation. Indeed, training subjects to identify such cognitive biases in their thinking, and to come up with strategies for avoiding them

[28] Freeman and Philippa A Garety 2004 argue that this is an important factor in the maintenance of persecutory delusions.

(metacognitive training), is an effective treatment for delusions (Moritz and Woodward 2007, Moritz, Andreou, et al. 2014). This indicates that such cognitive biases are a factor in delusion maintenance, functioning as masks on evidence-responsiveness capacities. And it indicates that these are not permanent masks, but can be removed with sufficient effort and motivation, allowing patients to successfully exercise their evidence-responsiveness capacities.

In sum, we have reason to think that subjects with delusions have the capacity to respond to counter-evidence on their delusion (subsection 3.3.2). This capacity can be masked by continued strange experiences, motivational factors, and cognitive biases, in which case the delusion stays in place, or does not sufficiently change, in the face of counter-evidence (subsection 3.3.3). However, when they employ that capacity in the absence of masking factors, the result is a rationally permissible change of mind in response to the evidence. This is what we see when subjects become less confident in the content of their delusion in the face of counter-evidence, when CBT succeeds, and in cases of evidence-responsive remission more generally.

## 3.4 Implications of the View that Delusions are Evidence-Responsive

### 3.4.1 Implications for the anti-responsiveness and anti-doxasticism arguments

The claim that delusions are evidence-responsive bears directly on the debate on the nature of delusions and belief outlined in section 3.2. As we saw there, both proponents of the Anti-Responsiveness and Anti-Doxasticism arguments hold that delusions are intractable, that is, not evidence-responsive, and rely on this claim as a central argumentative hinge. The discussion in section 3.3 shows that this view is false: delusions are evidence-responsive. We can therefore simultaneously hold the two following views:

> **The Evidence-Responsiveness View of Belief**: Belief is constitutively evidence-responsive, i.e. if $S$ believes that $p$, then $S$ has the capacity to rationally re-

spond to evidence bearing on $p$.

**Doxasticism**: Delusions are beliefs.

It follows that the Anti-Responsiveness Argument does not go through: delusions do not provide a counterexample to the claim that belief is constitutively evidence-responsive. Similarly, the Anti-Doxasticism Argument fails: the evidence-responsiveness of beliefs does not militate against the claim that delusions are beliefs. We can have our evidence-responsive cake and eat it too.

One might argue that this is too quick—for two reasons. The first has to do with the scope of the claim that delusions are evidence-responsive: to dissolve the debate I would need to show conclusively that all delusions are evidence-responsive, and I have not done so. The second has to do with the sense of evidence-responsiveness at stake: participants in the initial debate, this objector claims, meant evidence-responsiveness in a stronger sense. I will consider these in turn.

First, the scope objection. As I mentioned at the beginning of section 3.3, I cannot decisively show that *all* delusions are evidence-responsive, but only make an inductive case for that claim by considering data on a wide range of delusions. However, given delusional heterogeneity (in etiology, accompanying diagnoses, degree of circumscription, and content, among others), the worry remains that the results I present do not generalize to all delusions. Now, suppose that I have established that many or most delusions are evidence-responsive, but that there remain some that are not. Why not think that the very same debate arises again for this smaller subset of intractable delusions, with the Anti-Responsiveness side arguing that they provide a counterexample to the view that belief is constitutively evidence-responsive and the Anti-Doxasticism side arguing that they are not beliefs? And, if so, how does my argument in section 3.3 help with this debate?[29]

Against this, given that I present data supporting the evidence-responsiveness of delusions with a wide range of different contents, degrees, and kinds of circumscription, and

[29] Thanks to Christopher Willard-Kyle for pressing me on this point.

accompanied by a range of different psychiatric diagnoses, I think that we can legitimately conclude (through an inductive argument) that all delusions are evidence-responsive. Delusional heterogeneity in actual evidence-responsiveness is the result of different masks on evidence-responsiveness capacities.

Of course, the claim that all delusions are evidence-responsive relies on an inductive argument, so there may be delusions to which these findings do not generalize. But the burden of proof is on the opponent to present such cases. In other words, it is not clear if there is space left for the same debate to arise over a smaller subset of delusions. There is, to be sure, space for debates about how to classify imaginary cases, and to use thought experiments to probe the limits of belief. But this is a different debate, one that runs parallel to debates on delusions. This paper is not meant to refute every possible counter-example to the evidence-responsiveness view of belief resulting from thought experiments.

But let's grant the objector that there are some intractable delusions. It is not clear that this would generate the same debate we saw. This is because few if any doxasticists hold that *all* delusions are beliefs: for instance, as Elisabeth Pacherie and Tim Bayne note, they "certainly do not intend to argue that all delusional states are beliefs" (T. J. Bayne and Pacherie 2005, p. 179). For this reason, it is not clear that doxasticists would feel compelled to claim that these marginal cases are beliefs and use them to argue against the evidence-responsiveness view of belief. Indeed, the existing debate in the literature focuses precisely on the kinds of cases I consider in section 3.3, so it is fair to say that my discussion in that section advances the debate by showing that those are cases where we detect evidence-responsiveness (against the assumption made by both sides).

The second objection is that, in claiming that delusions are evidence-responsive in the capacities sense articulated in subsection 3.3.1, I have just changed the subject. Participants in the initial debate meant something else by "evidence-responsiveness", namely, rationally responding to counter-evidence most of the time, and it is no resolution of the debate to change the topic.

In response, grant that many participants in the debate spell out evidence-responsiveness along such lines. Most explicitly, Lisa Bortolotti's arguments (in Bortolotti 2005a, Bortolotti 2005b, and Bortolotti 2009) have as their target what she calls "the background view" of rationality constraints on belief, according to which belief requires an overall background of rationality. On the background view, "if deviations from norms of rationality are too widespread, then the ascription of beliefs is compromised" (Bortolotti 2005a, p. 20). Applying this view of rationality constraints to evidence-responsiveness, her target is the view that people rationally respond to counter-evidence on their beliefs most of the time.

I agree that we ought to reject this view. As Bortolotti 2009 notes, we do not even need to appeal to delusions to establish this point. We often fail to rationally revise many ordinary beliefs—including political beliefs, beliefs one cherishes, and prosaic beliefs about the goodness of our own everyday decisions—in the face of counter-evidence.[30]

However, this does not settle things in favor of the Anti-Responsiveness side. Proponents of the Anti-Responsiveness Argument take themselves to be attacking the most promising version of a connection between evidence-responsiveness and belief.[31] If I am right, then they are not attacking the most promising version of this idea. The version I propose is more promising, given that, unlike the ones they consider, it can accommodate data on delusion maintenance and provides a framework in which to theorize about the interaction between evidence-responsiveness, perceptual experience, motivational factors, and cognitive biases.

Similarly, proponents of the Anti-Doxasticism Argument aim to "protect the idea of essentially rational belief from attack;"(Currie and Jureidini 2001, p. 161), and not specifically the idea that beliefs adjust to the evidence in rational ways most of the time. The appeal to capacities for evidence-responsiveness offers a new tool to defend that general idea. The Evidence-Responsiveness View of Belief, spelled out in terms of capacities, makes

[30] See Mandelbaum 2019 for compelling defense of this claim.

[31] Indeed, sometimes they explicitly consider other versions of the idea. See (Bortolotti 2009, pp. 18–21).

belief essentially rational in the sense that it is part of the nature of belief that beliefs are underwritten by evidence-responsiveness capacities; in other words, that, in the absence of interfering factors, beliefs are rationally updated in response to counter-evidence. This is compatible with a range of powerful interfering factors to such updating, as we saw in subsection 3.3.3. For this reason, as I argued in section 3.3, belief being essentially rational in this sense is compatible with delusions being beliefs.

The notion of evidence-responsiveness I propose therefore allows us to make compatible two attractive and otherwise jointly untenable positions: the traditional view that there is some connection between believing and rationally responding to evidence, and the claim that delusions are beliefs. This dissolves the debate we saw in section 3.2, which hinges on the claim that delusions are not evidence-responsive. And it opens the door to holding that delusions are evidence-responsive beliefs.

As discussed in section 3.2, the doxastic conception of delusions has many advantages. Centrally, the doxastic view does justice to intuitive and clinical classifications of delusions as beliefs, respects the continuity between non-delusional beliefs and delusions, and accounts for the value of ascribing delusional beliefs in explaining patients' behavior. These advantages are preserved in a view that claims that delusions are *evidence-responsive* beliefs.

Though I favor the view that delusions are evidence-responsive *beliefs*, this can be seen as an optional add-on to my central claim in this paper, namely, that delusions are evidence-responsive. You can accept that claim while rejecting a doxastic analysis of delusions. Delusions might fail to be beliefs because they fail to meet other necessary conditions on belief, such as characteristic connections to action, inference, and affect. This is more than a bare possibility: there are standing debates on whether delusions meet such necessary conditions on belief.[32]

[32] My argument in this paper suggests a novel argumentative strategy for dealing with these debates. Instead of asking whether subjects with delusions act, infer, and feel in belief-characteristic ways most of the time, we should investigate whether subjects with delusions have the relevant capacities (e.g. for acting on a belief, drawing inferences from it, and experiencing corresponding emotional reactions). Focusing on

Independently of how such debates turn out, and of whether we ultimately ought to classify delusions as beliefs, showing that the Evidence-Responsiveness View of Belief does not rule out classifying delusions as beliefs is significant. It dissolves a long-standing debate in the literature. And it shows that the Evidence-Responsiveness View of Belief need not be overly rationalist, or require us to idealize beliefs. Evidence-responsiveness is compatible with the pervasive interference of non-rational factors, to an extent that can lead to delusion.

### 3.4.2 Implications for the study of delusion and belief maintenance

In section 3.3, I argued that delusions are underwritten by evidence-responsiveness capacities that can be masked in a range of ways, including by abnormal perceptual experiences, motivational factors, and cognitive biases. This involved surveying and piecing together a wide range of empirical literature on delusions. One can view the result as an an integrated framework on which to think about delusion maintenance. Instead of, for example, seeing views that focus on how delusions are responsive to evidence (Maher 1974) as competitors to views that center motivational influences (like traditional psychodynamic views), we can factor in contributions of these different factors into a single model.

This paper provides only a framework, not a full model that can make concrete predictions about the maintenance of specific delusions. As our empirical understanding of the contribution of different factors to delusion maintenance grows, we might be able to construct a detailed model, one which assigns weights to these different factors for specific delusions, helping us understand the causal mechanisms behind delusion maintenance and yielding recommendations for interventions that are likely to lead to delusion remission in specific cases.

capacities is a promising strategy for accommodating pervasive irrationality on the one hand, and theoretical and practical roles that belief is called upon to play on the other. Exploring this possibility is beyond the scope of this paper.

The framework I propose may also help us understand *belief* maintenance, not just delusion maintenance. If one embraces the orthodox view that belief is constitutively evidence-responsive, then beliefs are in the province of evidence-responsiveness capacities—as are delusions. Further, the masking factors to which I appealed to explain delusional evidence-resistance (non-veridical perception, motivational factors, and cognitive biases) also apply to beliefs. This suggests that beliefs and delusions are regulated by at least some of the same cognitive mechanisms.

Consequently, if my view of delusions is right, we can look to delusions to better understand beliefs, and vice-versa. This fits with the methodology of cognitive neuropsychiatry, which uses data from "people with acquired disorders of cognition to constrain, develop, and test theories of normal cognitive structures and processes" (A. M. A. Davies and M. Davies 2009, p. 288), and models of normal cognition to investigate the causes of such disorders.[33]

This approach has been successfully applied to the study of delusion formation. It underlies the account of Capgras formation discussed in subsection 3.3.3, which relies on comparing Capgras patients' responses to faces to a model of normal face recognition (Ellis and A. W. Young 1990). Building on such work, two-factor theorists of delusions (who think that there is an additional factor on top of perceptual abnormalities involved in delusion formation and maintenance) have approached delusion formation and maintenance through comparisons with our best models of such processes in subjects without delusions (Stone and A. W. Young 1997, M. Davies and Max Coltheart 2000, A. M. A. Davies and M. Davies 2009, Max Coltheart 2007, Max Coltheart et al. 2011). The framework that emerges from seeing delusions as evidence-responsive can be seen as a contribution to this research project.

[33] There is substantial unclarity about what normality consists in: is it the statistically normal case (here, of belief maintenance)? The case where the systems involved in belief maintenance meet their function, whatever that may be? The case that satisfies norms of rationality? This is an important unresolved methodological question in cognitive neuropsychiatry. I am here just assuming that there is some thin (non-moral) notion of normality that allows us to study the functioning and malfunctioning of cognitive mechanisms. Thanks to August Gorman for discussion.

### 3.4.3 Implications for treatment

The view that delusions are evidence-responsive in the capacities sense has implications for treatment. Specifically, it encourages us to take seriously the view that offering counter-evidence to the delusion—when conditions for getting the patient to exercise their evidence-responsiveness capacities are in place—is a promising path to delusion remission. In practical terms, the advice is to invest more in cognitive-behavioral therapy coupled with interventions that remove masking factors.

This fits current treatment practices. As seen in subsection 3.3.2, cognitive-behavioral therapy is recommended for delusions. Further, within cognitive-behavioral therapy, attention is given to ensuring that background conditions for the patient to respond to evidence are in place. For instance, cognitive-behavioral therapy includes building a relationship of trust between the patient and therapist, without which it is unlikely to be effective (Greene 2005). Trust is a condition for the patient to accept the evidence the therapist provides, and it reduces the likelihood of defensive reactions to it. In other words, it increases the likelihood that the patient will exercise their evidence-responsiveness capacities without the interference of motivational factors that manifest themselves in defensive reactions.[34]

There is growing recognition of the role of motivational factors in mediating the effectiveness of CBT. Specifically, when patients respond defensively to their delusions being directly challenged—a clear sign of the influence of motivational factors via cognitive dissonance—CBT is unlikely to succeed. In such cases, meta-cognitive therapy (discussed in subsection 3.3.3) is recommended. Meta-cognitive therapy takes a 'back-door' approach, not directly addressing the delusion but instead enabling the patient to do so at a later stage (Moritz and Woodward 2007), in a way that avoids motivational factors taking up such a large role in responding to the evidence. In addition to setting up the

[34] For more on the role of disturbances in trust and communication in the maintenance of delusions, see Fuchs 2015 and Fuchs 2020.

stage for interacting with evidence in a way that puts less weight on motivational factors, meta-cognitive therapy trains patients to avoid common reasoning biases. As such, it can be understood as removing masks of a non-motivational kind, namely, reasoning biases.

Finally, antipsychotic medication can be understood as operating on perceptual masks on the patient's evidence-responsiveness capacities. It primarily operates on the patient's dopaminergetic neurotransmission (Gardner et al. 2005). As seen in subsection 3.3.3, abnormal perceptual experiences caused by dopamine deregulation, which affects perceptual salience, are one of the factors that explain delusion maintenance in the face of counter-evidence. Antipsychotic medication operates on this masking factor through regulating dopamine transmission.

In addition to helping account for current treatment practices, the view that delusions are evidence-responsive helps de-stigmatize delusions, which is important for successful treatment. Delusions are not bizarre attitudes outside the space of reasons, or which place patients into a realm apart from people without psychosis. Instead, they are the result of (heightened forms of) cognitive, motivational, and perceptual factors that also cause and maintain ordinary beliefs. Normalizing and reducing stigma increases treatment motivation in patients (Lüllmann and T. M. Lincoln 2013). My view suggests that de-stigmatizing delusions is not merely practically helpful in treatment, but is also accurate to the cognitive mechanisms behind delusions.

### 3.4.4 Ethical implications

As the point about stigma foreshadows, the view that delusions are evidence-responsive beliefs has important ethical upshots. Delusions are often treated as grounds on which to exclude people from the moral and epistemic community. The view that delusions are evidence-responsive offers tools to argue that such exclusion is inappropriate, and not just for moral reasons.

The justification for excluding people with delusions from the moral and epistemic community goes something like this. People with delusions are unmoored from shared reality, beyond the reach of reasons or evidence. At least insofar as the delusion is concerned, there is no point attempting to rationally engage with them. Further, the strangeness and tenacity of many delusions easily make it easy to slip into thinking that there is something so deeply wrong that attempting to rationally engage on *any* topic is unwarranted. As a result, a kind of blanket objectifying treatment comes to cover all interactions with the patient.

In this view, a person with delusions is properly treated in *the objective stance*, as an object to be handled as opposed to a person to reason with. Indeed, people with delusions have been taken to be *paradigmatic* candidates for the objective stance. As Strawson noted in introducing the concept of the objective stance, seeing an agent "as one whose picture of the world is an insane delusion...tends to inhibit ordinary interpersonal attitudes in general, and the kind of demand and expectation which those attitudes involve; and tends to promote instead the purely objective view of the agent as one posing problems simply of intellectual understanding, management, and control"(Strawson 1962, pp. 16–17).

Now, one may think the objective stance is appropriate given the kind of attitude delusions are, yet oppose taking it up based on ethical considerations. For example, one may object to this treatment because it is disrespectful, or because marginalizing and stigmatizing people with delusions makes their lives worse and reduces the chances of delusion remission, while holding that people with delusions are outside the space of reasons, at least insofar as the delusion is concerned. This position makes adopting a non-objective stance toward people with delusions a useful fiction.

In contrast, my view of delusions implies that it is taking up the objective stance that is grounded in a (pernicious) fiction. It is not that it is morally better to act as if people with delusions are within the reach of reasons. They genuinely are within the reach of reasons. In other words, my view reconstitutes how it is appropriate to look at people with delusions, and opens up space for attitudes "of involvement or participation in a human relationship"(Strawson 1962, p. 9).

There is substantive debate about what these attitudes look like in the case of severe mental illness. The traditional Strawsonian view upholds a dichotomy between the objective stance and the participant stance, where the participant stance includes blame and resentment (in addition to attitudes such as "gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other"(Strawson 1962, p. 9)). However, perhaps blame and resentment are inappropriate in the case of delusions. Philosophers of psychiatry have challenged the claim that it is appropriate to blame patients while rejecting the view that we ought to shift into the objective stance. This can be achieved in two main ways: either by severing the participant stance and the appropriateness of blame (Pickard and Ward 2013, Pickard 2015) or by introducing new non-objective stances, such as "the nurturing stance" (Brandenburg 2018).[35]

Outlining precisely what attitudes are appropriate in the case of delusions is beyond the scope of this paper. The key point for my purposes is that the objective stance is not appropriate given the nature of the delusion, against the standard view outlined above. In particular, at least some attitudes and ways of interacting that fall sharply outside the objective stance, such as reasoning with someone and holding them to normative standards for how they ought to respond, are appropriate.

The fact that there is space for such attitudes does not on its own imply that we ought, all things considered, to take them up in every interaction. This is a familiar point. Taking up the objective stance is "a resource" we use "as a refuge…from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity"(Strawson 1962, pp. 9–10). This is true in interacting with anyone, not just with people with delusions. In the latter case, the strains of involvement may prove themselves more burdensome, so that taking up the objective stance when interacting on the topic of the delusion may still often be the better option.

[35] Thanks to Sofia Jeppsson for bringing the nurturing stance to my attention.

As the point above illustrates, the question of what stance to take up when interacting with people with delusions in specific circumstances is a difficult one. Addressing it would require taking into account many factors beyond the evidence-responsiveness of the delusion. Nevertheless, establishing that there is room for taking up a non-objective stance when interacting on the topic of the delusion is significant, especially given the marginalization, exclusion, and stigmatization that comes with the objective stance.[36]

### 3.5 Conclusion

Delusions have been taken to pose a hard challenge for the orthodox view that belief is constitutively evidence-responsive. They are typically classified as beliefs, yet they are deeply evidence-resistant, seeming to constitute a counterexample to the orthodox view. In the opposite direction, proponents of the orthodox view have argued that the deep evidence-resistance of delusions implies that they cannot be beliefs.

Against this, I appealed to empirical evidence on delusions and their remission to argue that delusions are evidence-responsive in the sense that people with delusions have the capacity to rationally respond to evidence bearing on their delusions. Their extreme evidence-resistance is a consequence of masking factors on these capacities such as strange perceptual experiences, motivational factors, and cognitive biases.

Once we see that delusions are evidence-responsive, we can dissolve the apparent tension between the view that delusions are beliefs and the evidence-responsiveness view of belief. Further, the claim that delusions are evidence-responsive yields a unified framework on which to study delusions and belief updating, and has significant implications both for treatment and for how to interact with patients with delusions.

[36] Thanks to August Gorman for helpful discussion of the ethical implications of the view.

# CHAPTER 4

# BELIEF CHANGE & SOCIAL CHANGE

**Abstract:** Individual beliefs often push back against structural reform, posing an obstacle to social change. Troublingly, beliefs that play this role are hard to change, often resisting counter-evidence. This poses a problem for structuralism, which prescribes structural change without considering how to get individuals to abandon resistant social beliefs. I argue that the structuralist has resources to address the problem of resistant social beliefs. Specifically, I argue that social network change can lead to the abandonment of resistant social beliefs, addressing even forms of active psychological resistance to belief change such as identity-protective reasoning. This solution to the problem of resistant social beliefs has significant implications for the debate between structuralists and individualists. In particular, it shows that careful attention to human psychology and proposing structural interventions are compatible. This makes room for bringing together insights from both individualist and structuralist traditions, allowing for a unified account of the relationship between belief change and social change.

> And there's no point sidling up crabwise with a mea culpa look, insisting it's a matter of the salvation of the soul. Genuine disalienation will have been achieved only when things, in the most materialist sense, have resumed their rightful place.
>
> Franz Fanon, *Black Skin, White Masks*, xv

## 4.1 Introduction

People with a criminal record find it hard to get jobs. And, when unemployed, people with a criminal record are more likely to re-offend and end up back in prison. Given that a disproportionate number of incarcerated people in the US are Black and Latinx, this vicious cycle contributes to racial injustice. To address this, activists have advocated for 'Ban the Box' measures, which make it illegal for employers to ask about criminal record in job application forms (Avery and Lu 2021).[1]

This is a structural intervention. Through a change in the law, 'Ban the Box' measures change the context in which particular individuals (in this case, employers) make decisions. Structural interventions contrast with individualist interventions, which aim to change individual attitudes as a way to change society. For example, trying to persuade employers to be less suspicious of people with criminal records is an individualist intervention.[2]

Intuitively, the structural intervention seems much more promising than individualist interventions in this case. Persuading employers one-by-one to be less suspicious of formerly incarcerated people is a tall order. In contrast, one would think that employers simply can't discriminate along these lines if they lack information on criminal records. Unfortunately, 'Ban the Box' measures do not seem to help—at least, insofar as the point of such measures is addressing racial injustice. Low-skilled Black and Latino men are marginally less likely to be employed after 'Ban the Box' measures than before (Doleac and Hansen 2016). When they are not allowed to ask about criminal record, employers' beliefs that Black and Latinx applicants generally have a sketchy background kick in and lead them to avoid hiring them. For this reason, in Michelle Alexander's words, "Banning the box is not enough. We must also get rid of the mind-set that puts black men 'in the box' " (Michelle Alexander 2010, p. 153).

[1] Madva 2020 discusses this example.

[2] See Ayala-López and Beeghly 2020, Brownstein et al. 2021, and Madva 2020 for more on how to draw this distinction.

As this case illustrates, individual attitudes—in particular, individual beliefs—can push back against structural change. In the 'Ban the Box' case, individual employers' beliefs lead them to find workarounds to keep up the (racist) *status quo*. In other cases, structural measures lead to individual complacency and moral licensing, with individuals assuming that a few measures mean that fairness has been achieved (Dover et al. 2014, Kaiser et al. 2013). Even worse, we sometimes encounter aggressive individual backlash to structural measures, as has sometimes happened with affirmative action (Hughey 2014).

Given that individual beliefs push back against structural change, achieving social change requires changing beliefs (Madva 2016).[3] This is a notoriously difficult task. As Charles Mills vividly put it about white ignorance, many such beliefs are ones that "resist, fight back…[are] militant, aggressive, not to be intimidated, active, dynamic, refuse to go quietly" (Mills 2007, p. 13). As a result, even if agents receive counter-evidence to such beliefs, they are not likely to abandon them.

Achieving social change, then, requires us to contend with the problem of *resistant social beliefs*: beliefs that (a) pose obstacles to the success of structural reforms, and (b) actively resist counter-evidence. Such beliefs generate practical difficulties in achieving social change. In particular, they pose obstacles for structuralist proposals: it is not easy to see how a focus on structural reform can succeed in the face of resistant social beliefs.

In this paper, I will argue that the structuralist has resources to address the problem of resistant social beliefs, and propose a family of interventions that might help us address the problem in practice. Specifically, I will argue that social network change—in the form of making room for dispersed social networks and promoting strong social movements—is a powerful structuralist resource for getting agents to abandon resistant social beliefs. This provides a novel defense of the power of structural interventions: they can address even resistant social beliefs.

[3] Two clarifications. First, though the paradigmatic cases of social change in this literature have to do with addressing structural injustice, any kind of change in our social organization is covered by the debate. Second, beliefs are not the only aspect of our mental lives that can push back against structural change. We will need additional interventions to address other aspects.

Importantly, this defense of structural interventions is *methodologically* individualist, in that it will involve detailed consideration of psychological mechanisms involved in belief maintenance and revision. In this way, this paper pushes back against the tendency to view individualism and structuralism as all-encompassing frames, where one must pick a side and stick to it both in methodology and practical recommendations.[4]

Recommending structural interventions is usually accompanied by hostility towards attending to the nuts and bolts of human psychology (E. Anderson 2010, Frye 1983, Dixon et al. 2012, Haslanger 2015, Haslanger 2022a, Táíwò 2017). Conversely, addressing the contribution of psychological phenomena to social problems is typically taken to require individualist interventions—interventions that directly target individual attitudes, instead of the context in which those attitudes are produced and maintained (Garcia 1996, Stanley 2015). My argument suggests that such polarization between individualism and structuralism is misguided, opening new avenues for exploring and achieving social change. Practical structuralists should abandon their hostility to methodological individualism, and methodological individualists should broaden their sights to consider structural interventions.[5]

I will proceed as follows. In section 4.2, I will argue that the psychological mechanism of identity-protective reasoning plays a crucial role in sustaining resistant social beliefs. Addressing identity-protective reasoning has often been taken to fall squarely in the province of practical individualism. Against this, in section 4.3, I will argue that social-network-shaping interventions can effectively address identity-protective reasoning. In section 4.4, I will argue that such structural interventions are also powerful when it comes

[4] See Haslanger 2020 for the distinction between methodological individualism and structuralism. See also Ayala-López and Beeghly 2020, Brownstein et al. 2021, L. J. Davidson and D. Kelly 2020, and Madva 2016 for discussions of the distinction. Note that I use the term 'practical individualism/structuralism' where these other theorists use the term 'individualism/structuralism about interventions.'

[5] If you are a structuralist who is not convinced that that individual attitudes push back against social change, you may want to resist the suggestion that changing individual beliefs matters *at all* to social change. In that case, I suggest taking this paper as offering an argument against the practical individualist starting from methodological premises that they would accept. The idea is this: even if, like individualists do, you think belief change is important, you should still prioritize structural interventions (in a sense I will clarify in section 4.5).

to addressing other factors that sustain resistant social beliefs. Altogether, this supports the view that practical structuralism can address the problem of resistant social beliefs. The upshot is that methodological individualism is compatible with, and can even support, practical structuralism. This makes room for a novel position in the structuralism–individual-ism debate, combining careful attention to psychology with promoting structural interventions (section 4.5).

## 4.2 Identity-Protective Reasoning

To see how the practical structuralist might be able to address the problem of resistant social beliefs, I will start by focusing on identity-protective reasoning happens when agents attend, interpret, and respond to evidence in ways that enable them to protect cherished social identities (Kahan 2012, Kahan 2015, Kahan 2017).

I am starting with identity-protective reasoning for two reasons. First, as I will argue, it is a crucial factor behind a large number of resistant social beliefs. Second, addressing identity-protective reasoning appears like a prime candidate for individualist interventions, and therefore it presents an especially thorny aspect of the problem of resistant social beliefs for structuralists. It is commonly thought that structuralism has resources to produce belief change in cases where beliefs transparently mirror the world: structural interventions change what such beliefs mirror. But it is harder to see how structural interventions can address beliefs that are psychologically encased in defensive mechanisms (such as identity-protective reasoning) that protect them from reflecting changes in external circumstances. Indeed, Alex Madva has argued that practical structuralism relies on "an unduly passive view of hearts and minds"(Madva 2016, p. 716), one which ignores that our beliefs are often protected against the evidence in these ways. Consequently, showing that identity-protective reasoning can be addressed through structural interventions substantially bolsters the case for practical structuralism. Before doing so in section 4.3, we need to get clear on the psychology of identity-protective reasoning. That is the task of this section.

### 4.2.1 From social identity to identity-protective reasoning

Identity-protective reasoning is all about managing one's beliefs in ways that protect one's social identities. Social identities (including race, gender, class, religion, political alignment, sports fandom, and so on) are part of our self-concept (i.e individuals' sense of who they are; Brown 2000, Tajfel et al. 1979, Tajfel 1982). The social identities that figure in our self-concept are thick with descriptive and normative content. When, say, being a woman is part of someone's self-concept, this does not just mean that she would place herself in the extension of the concept WOMAN. We have conceptions or characterizations (Camp 2015) of our social identities. These include both claims about characteristic features of members of the corresponding kind or group, and claims about the norms that apply to members of the group (Knobe et al. 2013), sometimes including norms on what beliefs one should have. To a large extent, we pick up these claims from our culture, but different individuals will have different conceptions of the same social identity.[6]

As self-affirmation theory (Gilbert 2006, Mandelbaum 2019, D. K. Sherman and G. L. Cohen 2006, Steele 1988) documents, we strive to defend our self-concept. In other words, we want to have a stable and good self-concept.[7] More specifically, we desire to hold on to the view that the features we incorporate in our self-concept are valuable, and to preserve the same features as part of our self-concept.[8] Because the self-concept includes social identities, we strive to defend our social identities. Individuals desire to hold to the claim that e.g. being a man, an immigrant, or a Democrat (with all that the more specific features that this involves for them) is good and valuable, and that they truly belong to

[6] For more on what centering social identities in our thinking and action involves, see Camp and Flores 2022.

[7] The desires at play in this context are typically not conscious.

[8] There are some exceptions to this. People with severe depression (among other mental health conditions) typically have a *negative* self-concept (Tarlow and Haaga 1996). In this case, they don't attempt to defend the goodness of their self-concept (because it is not a positive one), but they do seek to defend its stability (Swann Jr 1992).

these categories.[9]

The desire to regard our social identities as stable and good, in turn, affects how we manage beliefs connected with these identities, and how we respond to evidence that threatens these beliefs. In other words, we engage in motivated reasoning when it comes to such beliefs: desires to protect the self-concept causally influence how we interact with evidence (Kunda 1990), affecting what we infer from the evidence we have, what evidence is salient to us, and what evidence we gather. By affecting how we interact with evidence, our desires affect what we end up believing. If we lacked these desires, we would interact with evidence differently, and in many cases we would end up with different beliefs.

Summing up, we cherish certain social identities—for example, partisan, racial, national, or professional identities. They become part of our self-concept, of our sense of who we are and of our value. In normal circumstances, we aim to defend both the goodness and stability of our self-concept. Part of protecting our self-concept is defending cherished social identities. And this, in turn, leads us to engage in motivated reasoning when it comes to beliefs that are tied to our social identities. Specifically, we put in effort to find ways to maintain these beliefs in the face of counter-evidence.

### 4.2.2 The long reach of identity-protective reasoning

The psychological processes outlined above most transparently show up when beliefs in the goodness of cherished social identities are under attack. For instance, they are liable to be at play when white people are accused of benefiting from racist social structures, an accusation which compromises connections between whiteness and innocence (Sullivan 2006). Given the important role of beliefs in the goodness of such social identities in producing resistance to progressive social change, this would be enough to at least make

[9] Other psychological factors conspire to make us strive to defend cherished social identities. For example, desires that those who are not 'on our team' be less competent and deserving—as a byproduct of the desire that 'our side' deservingly wins—can play a role (Klein and Kunda 1992). And there is some experimental evidence that our need to cope with mortality and meaninglessness leads us to reaffirm the distinctive value of our communities (Pyszczynski, Solomon, et al. 2015). See Quilty-Dunn 2020 for discussion of the multiplicity of sources of defensive reasoning.

identity-protective reasoning a barrier to such social change.[10]

But identity-protective reasoning extends much further: we have reason to think it causally sustains many resistant social beliefs that are not directly connected with social identities. For one, a key mechanism by which we support our sense of the goodness of our social identities is by making positive comparisons with salient outgroups (Brown 2000, Tajfel et al. 1979). For this reason, identity-protective reasoning can turn into identity-*attacking* reasoning: reasoning that contributes to holding negative beliefs about salient outgroups.

More strongly, identity-protective reasoning extends to beliefs that at first sight appear distant from social identities. Here is an example.[11] Meat-eaters often endorse "the four Ns" about meat-eating: they believe that eating meat is (1) necessary (e.g., for protein), (2) natural (i.e., humans are meant to do it), (3) normal (i.e., humans generally do it), and (4) nice (i.e., meat tastes good) (Piazza et al. 2015). It is hard to get meat-eaters to abandon these beliefs. Interestingly, men are less likely than women to abandon these beliefs in the face of evidence, and more likely to respond to arguments against them with defensive reasoning. A promising explanation for this finding is that this defensiveness is a manifestation of identity-protective reasoning. Given culturally dominant connections between meat-eating and masculinity ("Real men eat meat"), men are more likely than other people to incorporate meat-eating into their identity (Rothgerber 2013). As a result, arguments against the four Ns are more likely to be met with defensive reasoning. The desire to defend masculinity under a dominant cultural conception makes (many) men closed-minded about meat-eating.

A more general example of the long-reach of identity-protective reasoning comes from thinking about political partisanship. Defending partisan identities (Van Bavel and Pereira 2018) has the power to affect many of our beliefs. This is because partisan identities are

[10] The same point applies to non-progressive social change projects. Beliefs in the goodness and value of queer identities pose an obstacle to a society that uniformly includes traditional family structures.

[11] See Quilty-Dunn 2020 for discussion of this example.

typically associated with a wide range of empirical beliefs. For example, under some current conceptions, being a Republican is strongly associated with believing that climate change isn't real, that vaccines are unsafe, and that creationism is true (Rutjens et al. 2018). Indeed, an increasingly wide range of views, including *prima facie* non-political views (e.g. about coffee consumption or fashion), have come to be associated with partisanship (Dellaposta 2020). This means that attempting to protect one's partisan identity often results in defensiveness about very large swathes of a person's web of belief.

Finally, identity-protective reasoning is long-reaching in that it affects attitudes that are not beliefs.[12] This discussion does not cover only on-off beliefs: it also includes credences and suspension.[13] For simplicity, I will use the term 'belief change' to cover abandoning one's belief that $p$, ceasing to suspend on whether $p$, changing one's degree of belief in $p$, or coming to believe that $p$. Including suspension makes this discussion straightforwardly relevant to addressing active ignorance (Alcoff 2007, Medina 2013, Mills 2007). Further, if you think, as I do, that implicit biases are beliefs, my discussion will be relevant for addressing implicit bias.[14]

In sum, as individualists have emphashized identity-protective reasoning makes a significant contribution to maintaining doxastic attitudes in the face of counter-evidence, including (and perhaps especially) ones which pose an obstacle to social change. Specifically, I have argued that we have good empirical reason to think that identity-protective reasoning is at play for many of our socially relevant doxastic attitudes, leading us to maintain such attitudes in the face of counter-evidence. For this reason, we need strategies for addressing identity-protective reasoning to solve the problem of resistant social beliefs.[15]

[12] Thanks to Caroline von Klemperer for encouraging me to expand the scope of the account.

[13] At least certain types; see McGrath 2020 on varieties of suspension.

[14] See Mandelbaum 2016 for a compelling case for the claim that implicit biases are beliefs.

[15] Note that this does not mean that identity-protective reasoning *always* poses obstacles to social change, only that it sometimes does. Further, the epistemic and practical rationality of identity-protective reasoning in different contexts are topics that deserve careful analysis (see Kahan 2015 for one proposal), and which interact in interesting ways with debates about externalism and internalism in epistemology (Srinivasan 2020). For the purposes of this paper, all we need is the claim that some instances of identity-protective

### 4.2.3 Moderating factors for identity-protective reasoning

It is not the case that the mere fact of valuing a social identity *inevitably* leads to identity-protective reasoning on *any* topic related to that identity in *any* context.[16] Identity-protective reasoning is subject to a range of moderating factors, the manipulation of which modulates how much agents engage in it. This gives us a number of levers to address identity-protective reasoning.

First, the threat to an agent's self-concept has to be significant enough for them to engage in identity-protective reasoning. Specifically, it must be the case that we sufficiently value that identity, that the belief targeted is sufficiently important to that identity, and that the counter-evidence is strong enough (Liberman and Chaiken 1992). (Later in the paper, I will focus on such levers on identity-protective reasoning.) Further, the more secure our sense of self is, the less likely we are to engage in identity-protective reasoning. Threats to one dimension of the self-concept won't make much of a dent in one's self-esteem, and one will have more resources to affirm other aspects of the self (C. H. Jordan et al. 2003, Steele et al. 1993). For this reason, one can resort to self-affirmations to increase openness to evidence (G. L. Cohen, Aronson, et al. 2000, Ditto et al. 1998, Moreno and Bodenhausen 1999, Reed and Aspinwall 1998, D. A. Sherman et al. 2000, D. K. Sherman and G. L. Cohen 2002).

Second, engaging in identity-protective reasoning is effortful, requiring access to working memory. Subjects under cognitive load (for example, subjects made to solve a memory-loading arithmetic problem) will not succeed at identity-protective reasoning (Ditto et al. 1998, Kahan 2012, Valdesolo and DeSteno 2008). Manipulating cognitive load can give us additional ways of controlling how likely agents are to engage in identity-protective reasoning.

Third, directional motivations—desires to arrive at specific conclusions—are pitted

reasoning contribute to the maintenance of resistant social beliefs.

[16] Thanks to Susanna Schellenberg for encouraging me to clarify this.

against accuracy motivations—desires to arrive at the truth, whatever that may be (Kunda 1990). Consequently, if subjects are sufficiently motivated to have true beliefs, they are less likely to engage in identity-protective reasoning. We can moderate identity-rotective reasoning by encouraging agents to get at the true answer: M. Prior and Lupia 2008 found that paying survey respondents when asking them political knowledge questions results in a higher proportion of correct answers, as opposed to answers that reflect the partisan line.

Finally, we only protect a cherished identity if that identity is salient to us in the context (G. L. Cohen, D. K. Sherman, et al. 2007). It follows that another venue for addressing identity-protective reasoning is changing the salience of social identities: for example, people are less likely to engage in identity-protective reasoning on topics related to one's country in the absence of nationalistic cues than in the presence of flags and uniforms (G. L. Cohen, D. K. Sherman, et al. 2007).

In my view, we should incorporate all these strategies in our repertoire for addressing identity-protective reasoning.[17] But I want to focus on a different point of intervention, one that lies at the causal root of identity-protective reasoning. and which, I will argue, is especially amenable to structural interventions: changing the structure of reasoners' self-concepts. This includes three kinds of changes to the structure of the self-concept: changing the *content* of cherished social identities; changing *which* social identities agents incorporate into their self-concept; and changing the (relative and absolute) importance of different social identities to their self-concept.

Why not focus on going even deeper into the roots of the phenomenon, and eliminating identity-protective reasoning *entirely*? The extreme version of this suggestion, on which we stop trying to defend our sense of self, is unlikely to work. What we know about human psychology suggests that the fact that we try to defend our sense of self is built into the structure of cognition. Indeed, Gilbert 2006 compellingly argues that a *psycholog-*

[17] We can understand proposals to resort to *civic rhetoric* (Stanley 2015) as involving incorporating all of these strategies into the speech to which we resort.

*ical immune system* that defends us from threats to the self is an integral part of human cognition. Further, the fact that one's self-concept includes social identities appears universal, with some theorists going as far as thinking that "groups constitute individuals" (I. M. Young 2014, p. 176), so that one's self-concept ineliminably includes social identities. What is changeable, in contrast, is the structure of the self-concept, in ways I will now discuss.[18]

## 4.3 Changing Social Networks as a Means to Changing Social Identities

At the beginning of section 4.2, I pointed out that individualists often appeal to identity-protective reasoning to argue that structural interventions are insufficient. Against the view that identity-protective reasoning provides a decisive obstacle to practical structuralism, I will in this section argue that there are valuable structuralist resources for addressing identity-protective reasoning. Specifically, I will argue that changes in social networks are a powerful means to change the structure of the self-concept in ways that reduce and redirect identity-protective reasoning.[19]

### 4.3.1 Social environments systematically shape social identities

Our social environment constrains and shapes our social identities. It systematically constrains which identities are available, what their content is, and which ones are socially appropriate. This is intuitive. Think, for example, of how, in some environments, women are valued for centering motherhood in their self-conception, whereas in others doing so might be looked down upon; or of how queer communities delineate fine-grained gender and sexual orientation categories, which are not available in dominant heteronormative

[18] A more moderate version of the 'eliminate identity-protective reasoning' proposal (suggested to me by Susanna Schellenberg) deserves investigation. On this version of the view, perhaps we could get people to have such a secure, stable, and multi-faceted sense of self that it would be nearly impossible for anything to be felt as a real attack to it. Indeed, some of the interventions I will propose in section 4.3 may sometimes have this result.

[19] I will then (in section 4.4) argue that such changes in social networks also reduce the power of other defensive mechanisms that frequently sustain resistant social beliefs.

contexts.

Indeed, according to self-categorization theory (J. C. Turner and Oakes 1986, J. C. Turner, Hogg, et al. 1987, J. C. Turner 2010, J. C. Turner and Reynolds 2011), whether social aspects of the self-concept, as opposed to individual ones (personal traits, values, and preferences), take a central place in one's behavior is a function of social context. Contexts where there are rigid social boundaries, status differences, and conflict between groups drive individuals to center social (as opposed to individual) identities in their self-concept (J. C. Turner, Hogg, et al. 1987). This suggests that reducing social boundaries and creating more egalitarian forms of social organization can reduce how much agents center social identities.

Specifically, we categorize collections of individuals as forming a group to the degree (*inter alia*) that the perceived differences between them are less than the perceived differences between them and other people (outgroups) in the relevant context of comparison (the meta-contrast principle (J. C. Turner and Oakes 1986)). If this is right, what we consider our ingroup, and, correspondingly, which identities we take up, is deeply dependent on the motley of people that happen to be around us, and the relationships between them.[20] Finally, social context shapes the *content* of the social categories at play. Who counts as a prototypical member of a group, and what traits are taken to be characteristic, depends on the contrast class at play (Hogg et al. 1990), i.e., on who is the outgroup in the context.

None of this is to say that which social identities we center is *determined* by the people around us. Different people can and do behave differently in the same environment, drifting towards different social groups and adopting different social identities. There is an interplay between individual agency and the environment, leaving individuals space to embrace or reject social identities that are put onto them.[21] For the purposes of defending

[20] The reference to *perceived* differences means that the categories that individuals bring to their understanding of the world, their motives, interests, etc, also matter to this kind of classification (J. C. Turner, Oakes, et al. 1994).

[21] See Haslanger 2022b for illuminating discussion of how social structures and agency relate.

the role of structural interventions in shaping the structure of agents' self-concepts, what matters is that there are some (defeasible and context-sensitive) generalizations about how individual social identities are causally affected by social context—which there seem to be, if self-categorization theory is along the right lines.

Offering further support to the idea that social identities at the individual level are shaped by social context, the *common ingroup identity model* (Gaertner et al. 2000) indicates that cross-group contact (contact among distinct social groups in conditions that foster respect and a sense of community) often produces new shared group identities among participants. These identities can fully replace pre-existing ones. For example, in successful company mergers, people stop identifying with the smaller company for which they used to work, and adopt a new identity associated with the larger company (Giessner et al. 2012). New identities can also sit alongside pre-existing identities: think here of how college students who feel a sense of school pride often adopt 'student at University X' as part of their sense of self, alongside existing religious, ethnic, or political affiliations (Dovidio et al. 1998).

The key point in the common ingroup identity model is that *cross-group contact changes the social identities we take up.*[22] It gets us to adopt new social identities and to change the ranking of importance of our different social identities. In doing so, we become less motivated to single-mindedly defend pre-existing identities. Perhaps, as in the company merger case, those identities have disappeared from our self-concept, in which case we no longer need to defend them. Or, as in the university student case, perhaps we now need to balance out defending different social identities. Alternatively, having a richer self-concept, we find attacks to any one of its dimensions less threatening.

[22] The common ingroup identity model was put forward as a way of explaining Allport's contact hypothesis (Allport 1954) that prejudice can be reduced by cross-group contact in good conditions. I am here repurposing the common ingroup identity model to think about addressing identity-protective reasoning (as opposed to explaining the modulation of affective dimensions of prejudice).

### 4.3.2 Interventions to reshape social identities

Given the discussion above, re-shaping social networks can have deep effects on the structure of the self-concept. I will now consider two specific ways of re-shaping social networks that can get individuals to systematically reduce their attachment to social identities that support resistant social beliefs.

*Dispersed social networks*

Consider a social network where individuals have a large number of diverse social ties, belonging to a number of different community spaces which they flexibly enter and exit. Someone who navigates very different social environments in their workplace, religious community, children's school district, and through their hobbies will have such a social network. Such a social network will often foster a more positive role for social identities in one's self-concept than a network where one belongs to a single, rigid, clearly delimited social group.[23]

In particular, such dispersed social networks provides many opportunities for cross-group contact. Correspondingly (given the common ingroup identity model), they make room for individuals to adopt new social identities. Having a self-concept that incorporates a rich diversity of social identities makes individuals less likely to single-mindedly defend any particular one. It generates the need to balance out defending different social identities and it reduces the overall threat posed by attacks to a specific dimension of the self. Further, individuals will find different social identities (and conceptions thereof) salient as they move across these different social contexts. This will provide contexts for

[23] Such dispersed social networks need not amount to fully integrated societies, in the sense of "*comprehensive* intergroup association on terms of equality…[which] requires the full inclusion and participation as equals of members of all races in *all* social domains" (E. Anderson 2010, p. 112) (my italics). The existence of cross-group contact under positive conditions is compatible with the existence of community spaces which are not open to all. It only requires the existence of *some* spaces where boundaries are broken. Given concerns about the costs of integration for marginalized groups (e.g., loss of access to services that cater to their specific needs and of distinctive cultural and sense-making spaces (Shelby 2017)), I think we should be careful when proposing integration as the way to promote helpful cross-group contact. Thanks to Dee Payton and Samia Hesni for discussion.

open-mindedness on specific topics. Perhaps at church, where religious identity is highly salient, a person is likely to defend traditionalist beliefs about gender roles; but, when talking with people at their mixed-gender running team, much less so. Finally, having a social network that is mostly constituted of a large number of weak social ties, as opposed to a small number of strong ones, is likely to make social identities recede in importance in comparison with personal characteristics (Mutz 2006). This may reduce how much people engage in identity-protective reasoning across the board.

Focusing on partisan social identities will make this dynamic clearer. Most partisans in the US now live in partisan bubbles (Bishop 2009, Levendusky 2009). They barely ever interact with supporters of the other party. As a result, they have ceased to have cross-cutting social identities (i.e., social identities that encompass both Democrats and Republicans). As Democrats and Republicans share fewer and fewer (non-political) identities, they grow more invested in defending their partisan identities grows. In an important sense, their entire sense of self stands and falls with their partisan identity.

If social sorting is the problem, cross-group contact might be the solution. In support of this suggestion, there is evidence that Democrats who have cross-cutting affiliations, thereby sharing important identities with Republicans (and vice-versa) are less likely to engage in partisan identity-protective reasoning (Mason 2018). If we want to reduce partisan identity-protective reasoning, building networks that make room for shared identities among Democrats and Republicans is a good idea.[24]

### *Social movements*

Social movements provide opportunities for contact for people of different social identities *who share a commitment to the same goal*, and for conversations that aim at building solidarity and community. Such conversations tend to result in our noticing commonalities and coming to identify under a common header. Because they function as sub-community

[24] Networks with a large number of weak social ties have additional benefits for awareness of rationales for opposing views and for political tolerance and support for civil liberties (Mutz 2006).

spaces that foster solidarity, social movements are exemplary sites for the kind of cross-group contact that produces shared identities among members of distinct social groups.[25]

Despite having that in common with dispersed social networks, social movements offer a different set of benefits when it comes to identity-protective reasoning. A social movement is a "sustained campaign of claim-making, using repeated performances that advertise the claim, based on organizations, networks, traditions, and solidarities that sustain these activities." (Tilly and Tarrow 2015, p. 11). Unlike having a social network consisting exclusively of weak social ties, participation in social movements tends to generate strong senses of affiliation and community. Further, social movements tend to generate novel strong social identities, and to lead participants to reshape their sense of self around new ones, as I will now detail.

Social movements often creatively produce new social identities around their shared goals. This can happen as a by-product of collective action, as we witness with labor organizer identities. Through participation in the labor movement, racially inclusive union identities come to the fore and white identities recede, leading union membership to reduce racial resentment among white workers and increase support for policies that benefit Black people (Frymer and Grumbach 2021).

Other times, social movements explicitly aim to produce or reshape social identities. For example, the Civil Rights movement explicitly aimed to produce a new Black American identity that did not incorporate white-enforced stereotypes. In Martin Luther King's words, one goal of the movement was to make room for "the new Negro," "a person with a new sense of dignity and destiny, with a new self-respect" (Martin Luther King 1956). In a different political direction, the NRA has had great political success largely by managing to purposefully cultivate "a distinct, politicized gun owner social identity" (Lacombe 2019, p. 1342).

Further, social movements sometimes attempt to produce alternative conceptions of

[25] Thanks to Peter Railton for highlighting the role of shared goals.

identities seemingly unrelated movement participation. For example, the vegan movement has invested in building new images of masculinity that do not incorporate meat-eating, by bringing attention to hyper-masculine vegan athletes and highlighting how veganism can be taken to express certain stereotypically masculine traits, such as emotional stoicism and protecting others (Greenebaum and Dexter 2018).

In producing and inculcating social identities, participation in social movements often leads to changes in participants' self-concepts (Kiecolt 2000). Participants might discard or add new social identities to their self-concept: for example, they might cease to identify as victims and start identifying as activists. They might come to center some social identities more than they did before, and come to devalue identities that were once central to them. In this way, upper-middle-class women who participated in the feminist movement in the 70s often began to think of themselves as feminists, and to identify less as housewives (Breinlinger and C. Kelly 2014). Participants may also come to reconceptualize identities in ways that match those of their activist community, thereby changing their self-conception. A man who becomes involved in the vegan movement might not only come to center a vegan social identity in his self-conception, but also come to reconceptualize masculinity. Finally, if the social movement gets enough public attention, these identities have a chance to turn into culturally mainstream identities (or ways of conceptualizing a given identity), as has happened with feminist identities.[26]

In sum, involvement in social movements can powerfully reshape identities by leading participants to adopt or center (1) identities as members of that social movement, (2) other movement-relevant identities, or identities conceptualized as the social movement proposes, and (3) identities that are inclusive of people who were previously exclusively outgroup members. This will re-orient identity-protective reasoning in important ways. Defending masculinity under a conception where it does not involve meat-eating is compatible with being open-minded on evidence about the health of meat-eating or

[26] See Amenta and Polletta 2019 for an overview of the effects of social movements on broader culture.

the environmental effects of meat consumption. Ceasing to primarily defend whiteness, and instead defending one's identity as a union organizer, makes one more open-minded when it comes to racial injustice (including factual questions about American history or the racial distribution of wealth in America)—especially if this new identity is racially inclusive.

### 4.3.3 Social network change as a structural intervention

I have suggested two ways of reshaping social networks that systematically contribute to the restructuring of self-concepts. What the ideal shape for a social network is for the purposes of avoiding resistant social beliefs depends on the beliefs and social identities at play. Is the identity-protective reasoning we want to address most effectively combated by developing and coming to center strong new alternative identities? Or is the best strategy to reduce attachment to social identities in general or make room for more inclusive social identities? Depending on the answer, we might want to prioritize developing social movements (in the first case) or dispersed social networks (in the second case). Beyond that, of course, there is work to be done on what the overall shape of a society-wide social network should look like—narrowly, from the point of view of reducing pernicious identity-protective reasoning, and broadly, from the point of view of justice.

The key point for my purposes is the following: social network change can reduce identity-protective reasoning, making individuals open to changing their minds on socially significant topics. In this way, social network change can play a key role in solving the problem of resistant social beliefs. And, importantly, social network change is a form of structural change: a large-scale change in the context in which beliefs are formed and maintained.

There is wide consensus in the literature in classifying social network change as a form of structural change. O'Connor and Weatherall 2019, Madva 2020 explicitly list it as a structural intervention that contrasts with individual-level debiasing, and E. Anderson

2010's integrationist proposals (a form of social network change) are generally taken to be structuralist.

One might object that changes in social networks also involve changes in individual attitudes: individuals must come to enjoy spending time with a different set of people, and they might need to come to know their way about different neighbourhoods, learn to appreciate different ways of doing things, and so on. But this point applies to all sorts of reforms that we naturally describe as structural. For example, legal changes, or changes in the material set-up which individuals navigate, require specific agents to implement them, and, as such, involve individual attitudes.

To the extent that it is helpful to describe some interventions as structural and others as individual-level, changes in social networks are best construed as structural changes.[27] Specifically, they are not well-modeled by thinking of isolated individuals changing their minds one-by-one. Instead, they require large-scale and highly coordinated changes in practices, social norms, the material conditions in which individuals interact, and attitudes that sustain current social networks. As is characteristic of structural changes, non-coordinated individual action will have little impact (Haslanger 2022a).

An analogy will help see the structural nature of the changes I am suggesting. Consider the case of addressing psychiatric disorders. An individualist approach would prescribe a combination of medication, talk therapy, therapeutic "homework" exercises, and the like—without directly aiming to change the backdrop of the person's life, that is, their economic, material, and social conditions. In contrast, a structuralist approach would note the ways in which psychiatric disorders can be sustained by lack of life prospects, economic difficulties, social marginalization and isolation, among other structural factors, and aim to address such problems as a way of addressing such disorders.[28] The same

[27] Brownstein et al. 2021 argue against this distinction, proposing that it is a mere matter of framing. Of course, if you reject the distinction between practical individualism and structuralism, the question of the connection between methodological and practical positions in the individualism vs. structuralism debate does not arise. If you agree with Brownstein et al. 2021, the ey take-away from this paper is a concrete, empirically-informed proposal for a family of interventions that can address identity-protective reasoning.

[28] See Pickard 2020 for discussion of this distinction in the context of addiction treatment.

distinction translates over to the non-pathological case of identity-protective reasoning. Individualist interventions might propose information workshops or training sessions to help individuals avoid identity-protective reasoning (Madva 2017). These contrast with proposing changes in the social networks which individuals navigate day-to-day in their social interactions.

The central upshot of this section, then, is the following: against initial appearances, a form of structural change is a powerful lever for addressing identity-protective reasoning. Consequently, the practical structuralist has resources for changing even resistant, active beliefs that are protected against counter-evidence—the kinds of beliefs that have been claimed to be intractable for the practical structuralist.

## 4.4 The Deep Effects of Social Network Change on Psychology

Identity-protective reasoning is not the only factor that sustains resistant social beliefs. Fully addressing the problem of resistant social beliefs will require the practical structuralist to have resources to address other sources of the persistence of these beliefs.

Though it is beyond the scope of a single paper to consider how to address *all* sources of resistant social beliefs, I want to bolster the case for the role of structural interventions by showing that they can powerfully influence many central aspects of belief revision. Correspondingly, in this section, I will argue that social network change can also address other crucial factors that help maintain resistant social beliefs.

### 4.4.1 Other kinds of motivated reasoning

According to systems justification theory (Jost 2019), desires to justify the social *status quo* can influence our reasoning (much like identity-protective desires do). These desires arise because feeling good about the *status quo* increases satisfaction and reduces the uncertainty, threat, and social discord that would come from attempting to bring about social change (Jost 2019). And, much like identity-protective desires, these desires have long-ranging effects. They protect both beliefs that the current social system is good and beliefs that serve to justify it or make it appear natural (e.g. beliefs in the natural submissiveness of women or in the supposed meritocratic nature of our society). More generally, desires to justify the social *status quo* help maintain ideological beliefs (Haslanger 2011, Haslanger 2017, Shelby 2003).

Social movements, it turns out, can also address such desires. Given the source of these desires, reducing the felt uncertainty, threat, and social discord that would come from attempting to bring about social change, and providing sources of satisfaction that are compatible with disliking the *status quo*, will help reduce systems justification reasoning. Social movements can play a powerful role in both of these. In building community and making space for experiments in living (E. Anderson 2014), social movements provide new sources of satisfaction that are compatible with disliking the *status quo*. Indeed, given the oppositional nature of social movements, accessing these sources of satisfaction might actively involve negative attitudes towards the reigning order. Further, as a form of collective action, social movements make attempting to bring about social change less daunting (Haslanger 2022a). In these ways, they reduce the costs of questioning the *status quo* for individuals.

A second relevant set of desires behind resistant social beliefs are desires to have socially adaptive beliefs (D. Williams 2021), i.e., desires to have beliefs that we are socially rewarded for having and to avoid beliefs that we would be socially punished for having. Socially adaptive beliefs include beliefs that serve to signal allegiance to the groups to which we belong (e.g., beliefs about gun use when it comes to partisanship (Kahan 2012)); beliefs that are required for good standing in a group (e.g. beliefs about the literal truth of the Bible for many mainstream Christian groups); and beliefs that facilitate smooth participation in dominant cultural practices (e.g., beliefs in the beauty of small feet, which promoted participation in foot-binding in early 20th-century China (Mackie 1996, Sankaran 2020)). The desire to have socially normative beliefs leads people to interact with evidence in ways that promote maintaining those beliefs.[29]

The kinds of social network change that I discussed in subsection 4.3.2 can reshape the desires that sustain socially adaptive beliefs. It is a familiar point that social movements affect social norms (e.g., E. Anderson 2014, Bicchieri 2016, Haslanger 2015, Haslanger 2017, Haslanger 2019 Sankaran 2020). Different sub-communities develop and enforce different social norms, including social norms on belief. To the extent that individuals are immersed in different communities, they become responsive to different sets of social norms. In the context of the social movement, norms that pressured them to have a given belief in dominant contexts do not have a grip on them, making them open to counter-evidence.

Social movements also often play a key role in changing social norms writ large: a vocal minority opposing a norm can lead to wide reconsideration of the norm, and to the norm losing its authority (E. Anderson 2014). For example, public pledges by sub-communities played an important role in changing social norms about foot-binding in early 20th century China (Mackie 1996), including norms on *beliefs* about foot-binding. Such public pledges reduce the social costs of not following previously unquestioned norms. For example, the social costs to rejecting the (previously dominant) beliefs that foot-binding promotes good health and fertility were reduced as more and more people came to vocally reject such views.

Indeed, social movements can affect beliefs without any new evidence coming into the picture.[30] In developing and disseminating new social norms, social movements might make new beliefs socially normative, generating desires to have and maintain those new beliefs. For example, as campaigns against foot-binding succeeded, the belief that foot-binding is cruel and bad for one's health became socially normative, generating pressure

[29] There is some disagreement about whether such cases reflect real beliefs. Some theorists (Hannon 2021, Schaffner and Luks 2018) think that social pressure only motivates people to *appear* to have certain beliefs, which can be done without actual belief. For reasons that Quilty-Dunn 2020 and D. Williams 2021 point out, I think that at least some such cases amount to belief. If you disagree, you can just bracket this discussion, as the central point of the paper does not depend on it.

[30] Thanks to Hanna Pickard for discussion.

for individuals to have this belief. Similarly, as we saw at the end of section 4.3, social movements can change which social identities we defend, making us engage in identity-protective reasoning with respect to a different set of beliefs.[31]

In this way, belief change can be entirely driven by changes in social networks and norms and the changes in desires that these engender. This is an old thought: as Pascal 1852 put it, if you want to adopt certain beliefs, you should "Endeavour to convince yourself, not by increase of proofs, but by the abatement of your passions," where a crucial means to this is integrating yourself in communities that have the relevant beliefs. My discussion puts flesh to the bones of this point, by drawing on research in psychology to explain why Pascal's suggestion can succeed. Specifically, changing social communities changes many of our desires, which have to do with fitting in, having positive relationships with those closest to us, and feeling positive about those we see as 'our people'. Given that such desires profoundly affect belief revision, changing social communities can entirely re-orient out doxastic take on the world in a way that bypasses evidence.

Alternatively, social network change can genuinely make agents more open to evidence—not merely bump them from one belief to another. Much as social network change can make us engage less in pernicious forms of identity-protective reasoning, it can destabilize what beliefs count as normative without replacing them with new normative beliefs in the same domain. In this way, social network change can set the stage for agents to rationally respond to evidence, and thereby for genuine rational engagement.[32]

[31] I am not recommending this as a strategy for changing others' beliefs. In bypassing agents' epistemic rationality, this would be problematically manipulative. Deliberately implementing such a strategy risks eroding important democratic ideals of mutual respect and collective deliberation. My point is simply that changes in social networks can in and of themselves have deep effects on our beliefs. Insofar as we are interested in either understanding or promoting belief change in a social context, we need to be alert to this phenomenon.

[32] Here too one might have ethical worries about policy-makers devising and implementing social network change as a means to collective epistemic improvement, even if evidence and argument come in. Whether and when such interventions would infringe on our autonomy, and the limits of their moral permissibility, are open questions, and ones which would require engaging with the details of specific interventions. Thanks to Sophie Keeling for pressing this point.

### 4.4.2 Trust and access to evidence

Suppose that motivational factors have been cleared away and agents are open to evidence. This on its own does not suffice to get agents to abandon their resistant social beliefs. They must, additionally, have access to evidence that challenges those beliefs, and count it as evidence in the first place. For this reason, solving the resistant social beliefs problem requires good distribution of evidence and appropriate patterns of trust (of the kind that allow us to properly count others' good testimony as evidence).

Here, again, the shape of social networks turns out to be crucial. As recent work in social epistemology has emphasized, the fact that agents sometimes have biased samples of evidence and fail to trust reliable sources often has its roots in dysfunctional social networks. O'Connor and Weatherall 2019 persuasively argue that the shape of social networks affects the evidence agents have in ways that help explain the maintenance of false beliefs. C. Thi Nguyen 2020a argues that *echo chambers*, i.e. social epistemic structures which pervert and corrupt one's epistemic trust, pose additional problems that are independent of evidence access.

Restructuring social networks has a crucial role to play in addressing such problems. Opportunities for cross-group contact provide occasions for epistemic friction (Medina 2013), including for receiving counter-arguments to one's views. Such opportunities can also make individuals open to a wider range of sources of evidence. Group belonging affects trust. We tend to trust ingroup members and distrust outgroup members (Tajfel 1970), going as far as rejecting information offered by outgroup members while accepting the same information from ingroup members (J. C. Turner, Hogg, et al. 1987). To the extent that cross-group contact reshapes who counts as being in one's ingroup, they will reorient one's trust.[33]

[33] This observation does not provide a full solution to the echo chambers problem: as C. Thi Nguyen 2020a notes, it will be very difficult to get people to interact with outgroup members and to feel a sense of kinship with them if their starting point is demonizing outgroup members. The point is simply that, if we could get such interactions going, we should expect to see re-orientations of trust.

Reshaping social networks can also improve one's patterns of trust by reducing testimonial injustice. Testimonial injustice occurs when a hearer's identity prejudice leads them to ascribe less credibility than the speaker deserves (Fricker 2007). Reducing identity prejudice helps address testimonial injustice. To the extent that identity prejudice is constituted by prejudiced beliefs, and prejudiced beliefs are partly maintained in virtue of identity-protective reasoning, the interventions outlined in section 4.3 will contribute to addressing testimonial injustice.[34]

In addition to changing evidence access and patterns of trust, reshaping social networks can affect agents' perspectives (Camp 2017). As Fraser 2021 has compellingly argued, testimonial exchanges often do more than transmit evidence. They convey perspectives, i.e. suites of interlocking dispositions to attend, inquire, value, and interpret the world in specific ways (Camp 2017). To the extent that changing social networks changes whose testimony individuals hear, it expands our access to different perspectives. For this reason, cross-group contact can help us develop what Arendt 1989 calls an "enlarged mentality", characterized by having a wide range of standpoints present in one's mind and the imaginative capacity to occupy them.

In sum, social networks affect a wide range of *psychological* factors involved in the maintenance of resistant social beliefs. Given the importance of social networks in shaping our belief maintenance and revision—by affecting evidence access, trust, perspectives, and patterns of motivated reasoning—they should be at the front and center of our theorizing about socially significant belief change. This has implications for how we think of the interaction between structural and individual change. I turn to these in the next section.

[34] Fully establishing the role of social network change in addressing testimonial injustice would require determining to what extent identity prejudice is implemented in prejudiced beliefs, and how large the role of identity-protective reasoning is in the maintenance of prejudiced beliefs. If this line of reasoning succeeds, it offers further support structural interventions (E. Anderson 2012) as opposed to virtue cultivation (Fricker 2007, Madva 2019) to address testimonial injustice.

## 4.5 Integrating Methodological Individualism and Practical Structuralism

In the introduction, I noted that theorists of social change tend to split into two polarized camps: individualists and structuralists. Each side takes a distinctive party line on both methodology and practical recommendations. Individualists recommend detailed study of psychology and interventions that focus on changing hearts and minds one-by-one. At the other end of the spectrum, structuralists eschew the study of the mind and propose changes to large-scale social structures.

The discussion in this paper challenges this polarized field. We can and should integrate a measure of methodological individualism into our study of social change, and that doing so is compatible with, and may even support, practical structuralism.[35]

Let's start with practical structuralism. In this paper, I have argued that change in social networks is a powerful means for achieving change in individual beliefs. Social network change, I have argued in subsection 4.3.3, is a form of structural change. Therefore structural reform can be a powerful lever for changes in individual belief.

This is a novel point for practical structuralism. Structuralists have emphasized that some components of unjust social structures persist independently of the beliefs of (the vast majority of) individuals in that structure. Think here of the material set-up of buildings that are not accessible to people with physical disabilities or of complex bureaucracies. We need structural interventions to target such components. But, as we saw in section 4.1, we also need interventions targeting beliefs. Individualist-leaning theorists have taken this to checkmate structuralism. For instance, Madva 2020 writes that cases where our beliefs push back against structural change provide "an argument for insisting on the importance of *individual level* debiasing strategies, which change individual's biased assumptions" (Madva 2020, my italics).

Given my discussion in section 4.3 and section 4.4, this conclusion is unwarranted.

[35] Though this combination of views has been neglected in recent literature, it harks back to Fanon 2007's discussion of social change, which combines centering deep structural reform with attention to the deep psychological effects of oppression.

Structural interventions can play a key role in changing beliefs. To put it differently: prioritizing belief change is compatible, and may even support, prioritizing structural interventions..[36] By showing that structural interventions can powerfully target even resistant social beliefs—the cases that are supposed to pose the deepest problem for structuralism—the discussion in this paper substantially strengthens the case for practical structuralism.

I will turn, now, to considering methodological aspects of the structuralism–individualism debate. My discussion above pays close attention to the psychology of belief maintenance. This goes against recent trends in social epistemology, which explicitly reject appealing to psychological factors in explaining socially troubling beliefs. For example, O'Connor and Weatherall 2019 write that "to focus on individual psychology is to badly misdiagnose how false beliefs persist and spread" (O'Connor and Weatherall 2019, p. 7). Similarly, Thi Nguyen writes that we should not understand the persistence of socially pernicious beliefs "in terms of individual psychological tendencies, such as motivated reasoning" but of "systems and environments" (C Thi Nguyen 2021, p. 231).

As I discussed in section 4.4, these projects yield many important insights about the role of social networks in shaping evidence access and trust. However, eschewing appeal to psychological factors is misguided. As I argued in section 4.2, we have strong empirical reasons to think that identity-protective reasoning plays a role in the maintenance of resistant social beliefs. Ignoring its role leads to theoretical distortions in our understanding of the mind and of social change. It also leads to unpleasant practical surprises, as interventions that ignore identity-protective reasoning will encounter active resistance from believers. Without changes in desires and motivation, counter-evidence (from trusted sources) is likely to be resisted. For instance, when people are motivated to maintain a stereotypical belief, counter-evidence to that belief tends to result in sub-typing (i.e.,

[36] Two qualifications. First, accepting the power of structural interventions for belief change is compatible with also incorporating individualist interventions, such as individual debiasing (Madva 2017). Second, I am presupposing, as structuralists generally do, that there is a worthwhile distinction to be drawn between structural and individual interventions, even if structural interventions typically require individual action (Ayala-López and Beeghly 2020) and whether we are inclined to classify an intervention as structural is subject to framing effects (Brownstein et al. 2021).

coming to believe that the stereotype applies to a subset of the original group), not in the abandonment of the stereotype (Kunda and Oleson 1995, Richards and Hewstone 2001).

Further, against some structuralists' insistence on the irrelevance of psychology, *any* form of theorizing about belief change involves making assumptions about the structure of cognition. Despite their claim to eschew psychology, both O'Connor and Weatherall 2019 and C Thi Nguyen 2021 make assumptions about the psychology of individual agents: O'Connor and Weatherall 2019 assume a Bayesian cognitive architecture, and C Thi Nguyen 2021 assumes bounded rationality (i.e. ideal (Bayesian) rationality bounded by processing limitations).

The charitable way of reading these projects is as making use of idealization: they idealize away substantive deviations from simple rational models of cognition. Such idealization helpfully isolates the role of factors beyond such deviations in belief maintenance. However, it is important to keep in clear sight that we are idealizing, and to recognize that a complete explanation will have to bring in accurate psychological models. This paper contributes to this project, complementing the findings we arrive at through idealization.

Indeed, to the extent that structuralists rely exclusively on idealized models of cognition, they make themselves vulnerable to worries that "the policy predictions of structural prioritizers rely on oversimplified psychological models" (Madva 2016, p. 702). Specifically, Madva thinks that structuralism assumes a picture in which we get individual change for free from independently desirable structural reforms: *The Mirroring View of Beliefs*, i.e. the view that beliefs are "mirror-like reflections of local environments and communities within which individuals are immersed" (Dasgupta 2013, p. 240), reflecting the bad evidence that individuals have available to them. As Madva correctly points out, this view is false: individual minds actively resist counter-evidence in ways that allow us to maintain socially troubling beliefs.

It is true that structuralists have often endorsed something like the Mirroring View (e.g. Antony 2016, Huebner 2016). My discussion in this paper shows that practical structuralism does not rely on this view. Focusing on the ways in which we actively resist the evidence in fact *supports* structural interventions. Indeed, structuralists have independent reasons to be suspicious of the Mirroring View. Structuralists often suggest that our cognitive structures are partially the result of internalizing social structures (Zheng 2018). Internalizing social structures involves actively filtering the world in ways that express the effect of social norms and identities on our epistemic agency. It is open to the structuralist to think of the structural context for belief revision as including aspects of our minds that are deeply shaped by social structures, such as identity-protective reasoning. For these reasons, structuralists should be much friendlier to psychology than they currently tend to be.

## 4.6 Conclusion

In theorizing about social change, paying close attention to the psychology of belief change, and especially to "deviant" factors such as identity-protective reasoning, is typically taken to lead to all-encompassing individualism about social change. Against this, I have argued that structural change (in the form of social network change) can address identity-protective reasoning. Indeed, changing the shape of social networks can deeply restructure our epistemic relationship with the world, affecting the evidence we have available, who we trust, and a range of forms of motivated reasoning.

Altogether, this brings a new perspective to the debate between individualism and structuralism. Individualism (and structuralism) about methodology and interventions are separable. Methodological individualism need not support individualist interventions. Focusing on psychology can *support*, instead of undermining, structural interventions, and it is compatible with recognizing the deep effects of social structures on psychological structures. Noticing these points makes room for integrating important insights from both individualist and structuralist traditions. As structuralists suggest, it is important to attend to individuals *qua* nodes in social structures, and recognize how deeply structures shape individual minds. But, as individualists emphasize, doing so requires integrating realistic models of human psychology. In doing so, we can recognize the power of structural changes for changing beliefs without seeing beliefs as simple mirrors of social reality.

# CHAPTER 5
# EPISTEMIC STYLES

**Abstract:** Epistemic agents interact with evidence in different ways. This can cause trouble for mutual understanding and for our ability to rationally engage with others. Indeed, it can compromise democratic practices of deliberation. This paper explains these differences by appeal to a new notion: epistemic styles. Epistemic styles are ways of interacting with evidence that express unified sets of epistemic values, preferences, goals, and interests. The paper introduces the notion of epistemic styles and develops a systematic account of their nature. It then discusses the implications of epistemic styles for central questions in epistemology, in particular, for issues surrounding rational engagement and for the debate between virtue epistemologists and epistemic situationists.

## 5.1 Introduction

People interact with evidence in different ways. Evidence that persuades you might leave others cold or lead them to strengthen their views. What indicates nefarious intentions to one person suggests bumbling incompetence to another. Where one person briskly rules out alternative explanations, another keeps them alive, refusing to make up their mind.

This variation in ways of interacting with evidence compromises our ability to understand one another. And it poses problems for rational engagement, endangering democratic practices of collective deliberation. This makes it important to address why people interact with evidence in different ways.

In this paper, I discuss a neglected ingredient behind how people interact with evidence, one which plays a crucial role in explaining systematic differences in modes of epistemic engagement: *epistemic style*. Though the notion of epistemic style has remained under-theorized, the phenomenon is familiar. It is exemplified in the charge that American politics has come to be dominated by the " paranoid style," which expresses "heated exaggeration, suspiciousness, and conspiratorial fantasy" (Hofstadter 2012). Epistemic styles are also at play in the distinctive ways of interacting with evidence that some intellectual communities—such as Black feminists or the self-described rationalists—seek to articulate and inculcate.

I analyze epistemic styles as unified ways of interacting with evidence which express a cohesive set of epistemic parameters, and which agents can put on and take off. I argue that differences in epistemic style are at play in paradigmatic cases of systematic differences in how people interact with evidence.

This goes against standard views in the literature, which tend to account for such differences in two ways: either by appealing to epistemic virtues, vices or other deep character traits (in *virtue-theoretic approaches*), or by appealing to the effect of irrelevant contextual factors (according to *situationists*). Unlike virtue-theoretic approaches, my view does not impute deep, long-standing character traits to agents to explain their ways of interacting with evidence. For this reason, the view avoids concerns about the existence and explanatory power of such robust traits in ordinary agents. At the same time, unlike situationist approaches, my view does not portray agents as passive conduits for their context: their interactions with evidence remain the result of epistemic parameters of their own. For this reason, my account helps us address the long-standing debate between virtue theorists and situationists.

Further, this account of the ways in which agents interact with evidence provides us with tools for understanding others *qua* epistemic agents and for designing more effective strategies for rational engagement. And, because we can only begin to properly assess our interactions with evidence once we are clear on their roots, it functions as a prolegomenon to a novel approach to central questions about how to epistemically assess interactions with evidence.

The plan for the paper is as follows. In section 5.2, I articulate and motivate the central question of the paper and key desiderata for a good answer. In doing so, I argue that existing approaches to how people interact with evidence are insufficient. In section 5.3, I develop my analysis of epistemic styles. In section 5.4, I employ the notion of epistemic styles to put forward my explanation of why epistemic agents vary in how they interact with evidence, and show how this explanation meets the desiderata outlined in section 5.2. Finally, in section 5.5, I sketch implications of thinking of epistemic behavior in terms of epistemic styles for epistemology.

## 5.2 The Variation Question

Different people—and the same person in different contexts—interact with evidence in different ways. They update their attitudes differently in light of the same evidence, differ in the beliefs on which they take evidence to bear, explore different explanations for evidence, assess sources differently, and so on. And they inquire in varied ways in the same epistemic situation, differing in how they gather evidence, ask questions, and generate explanations. This raises the following question:

> **The Variation Question**: Why is there inter- and intra-personal variation in ways of interacting with evidence?

This is a descriptive question: it asks for an explanation of people's interactions with evidence. We can offer such an explanation without presupposing that those interactions are rational. I will, in my discussion, remain as neutral as possible on which ways of interacting with evidence are rational.[1] At the same time, the answer to the Variation Question has normative implications. Properly understanding what is going on at a cognitive level

[1] There is a substantial literature on this question, more specifically, on which doxastic adjustments in response to evidence are rational. On the side of 'there is precisely one rational adjustment' (the uniqueness thesis), see White 2013, Dogramaci and Horowitz 2016, Schultheis 2018. On the permissivist side, according to which there can be more than one rational response to evidence, see Douven 2009, T. Kelly 2013, Willard-Kyle 2017, Callahan 2021.

when people interact with evidence matters for assessing those interactions. I will discuss normative implications of my answer to the Variation Question at the end of this paper.

A central reason to be interested in the Variation Question comes from the social and political significance of the fact that people interact with evidence in different ways. First, this diversity makes trouble for mutual understanding. Encountering a person who interacts with evidence in ways that radically diverge from our own can be disconcerting, generating a sense of distance and alienation. Why would someone act *like that*? What could they possibly be thinking? This can easily lead to thinking that it is not worth engaging. And, when we engage less with others, the chances for understanding diminish. The result is a vicious loop where the prospects for mutual understanding continually thin down, and where mutual alienation and distrust continually poison the social waters.

This has important political consequences. As political scientist Michael Morrell notes, without mutual understanding,

> it is highly unlikely that citizens will demonstrate the toleration, mutual respect, reciprocity, and openness toward others vital for deliberative democracy to fulfil its promise of equal consideration that is central to giving collective decisions their legitimacy. (Morrell 2010, pp. 114–5)

In addition to posing problems for understanding, variation in ways of interacting with evidence poses problems for rational engagement. Without knowing how an agent will respond to evidence, how do you select evidence that will help you productively engage? Without a realistic shot at rationally engaging, the scope for joint deliberation becomes highly limited. This is a problem for democracy, which normatively relies on rationally persuading others and (on popular accounts) on collective deliberation (Dryzek 2002, Estlund 2009, Landemore 2017).

These problems motivate the need for an answer to the Variation Question. And they constrain the shape that such an answer should take: we want an account that can help us begin to address the issues I have just outlined. Such an account should meet the following two desiderata:

> **Prediction Desideratum**: To put us in a position to predict how others will interact with a range of evidence.
>
> **Understanding Desideratum**: To put us in a position to understand others' interactions with evidence.

Ideally, an answer to the Variation Question should help us predict how others will interact with evidence so that we can better select strategies for rational engagement. And it should help us understand—make rational sense of—why others interact with evidence in the ways they do.

These are only two of the desiderata that an answer to the Variation Question should meet. To outline a few more, I will now consider candidate answers to the question and why they fail.

An initially attractive idea is that the answer to the Variation Question is simple: *modulo* performance mistakes, people interact with evidence differently because they have different beliefs about the topic under discussion. If two epistemic agents interact with evidence differently, it is either because some of their beliefs about the topic at hand differ or because at least one of them made a reasoning mistake.

This view follows from (but does not require) two popular assumptions: (a) there is only one way to reason rationally once we fix beliefs and evidence (White 2013, Dogramaci and Horowitz 2016, Schultheis 2018) and (b) modulo performance mistakes, all epistemic agents reason rationally (e.g. D. Davidson 1985, Dennett 1981, D. Lewis 1974, Stalnaker 1984). It also fits with a natural construal of Bayesianism about human reasoning (Clark 2013, Friston 2012, Oaksford, Chater, et al. 2007, Tenenbaum et al. 2011). Bayesians think that (modulo performance mistakes) everyone reasons according to Bayes theorem. If we fix beliefs and evidence, then there is a unique response to evidence: the one that Bayes' theorem yields.[2]

Indeed, there are many cases where differences in beliefs about the topic or performance mistakes fully explain differences in interactions with evidence. For example, if two people look at a restaurant bill and come to different beliefs about how to split it, this is likely the result of a performance mistake or of different beliefs about who should pay for what. But not all cases fit this simple model.

My focus is on explaining what is going on in cases that do not fit this model. Such cases are commonplace. For one, subjects often set different *evidential thresholds* for revising their beliefs: whereas one person might change their mind on $p$ given very little counter-evidence, another might require a large quantity of counter-evidence to do so. As an example, gritty people set high evidential thresholds for abandoning the belief that they are likely to succeed (Paul and J. Morton 2018), meaning that it takes a lot of good evidence to persuade them that they are not likely to succeed. Defeatists do the opposite, taking even the smallest setback to show that they are doomed.

More generally, two agents starting from the same beliefs about the topic under discussion and the same evidence might nonetheless set different evidential thresholds on the same topics.[3] The gritty person and the defeatist might both start with the same belief that they are likely to succeed, and the same beliefs about what factors contribute to success, and still set different evidential thresholds.[4] As a consequence, one of them might

[2] This requires claiming that the likelihood function is encoded in beliefs. If one rejects this assumption, the view I will put forward is compatible with Bayesianism. To anticipate, Bayesians can see agents' epistemic styles as fixing likelihood functions.

[3] I am not making any claims about the epistemic permissibility of setting different evidential thresholds.

[4] One might object that differences in evidential thresholds must ultimately reduce either to performance mistakes or to differences in belief. Indeed, Paul and J. Morton 2018 hold that evidential policies are "implicit attitudes or guidelines" (Paul and J. Morton 2018, p. 191), which one can plausibly construe as beliefs.
However, even if epistemic parameter settings are all ultimately reducible to beliefs—an open question—the view we end up with by accommodating epistemic parameter settings in our model is very different from the naive view above. The view now is now that differences in interactions with evidence reduce to differences in (a) beliefs on the topic at hand, (b) performance mistakes, and (c) (probably implicit) beliefs about how to interact with evidence. Factor (c) is not in the naive view. Further, more research is needed to determine whether evidential threshold settings are implemented or determined beliefs: this would require, for example, determining whether one's evidential threshold settings are sensitive to evidence in a belief-like way and interact with desires in a belief-like way. We are better off focusing on the role of epistemic parameter settings in how agents interact with evidence, and leaving the question of implementation for

revise beliefs that the other does not in the light of the same evidence. In that respect, at least, they will interact with the evidence differently.

To accommodate the significance of differences in evidential thresholds, any answer to the Variation Question should meet the following desideratum:

> **Epistemic Parameters Desideratum**: To account for the fact that at least some differences in interactions with evidence are the result of differences in epistemic parameter settings.

A natural suggestion in light of the discussion above is that we can explain differences in how individuals interact with evidence in terms of a specific epistemic parameter: evidential thresholds across one's belief set (one's evidential policy).

However, there are causally relevant epistemic parameters beyond evidential policies. For example, agents differ in how much they value getting truth over avoiding falsehood (James 1979): if given the choice between acquiring 101 true beliefs and 100 false beliefs or acquiring no new beliefs, some agents will prefer the former and some the latter. People weigh theoretical values differently: the Quinean with a preference for desert landscapes will opt for a theory with few postulates, whereas the maximalist will prefer a complex theory that fits more data. As a result, they will come to different beliefs based on the same evidence. Further, agents find the same evidence compelling to different extents. For example, some people are unlikely to change their mind based on first-hand testimony, but find statistical surveys highly persuasive, whereas others have the opposite preference.

On top of this, it is unlikely that a single parameter can explain differences in *all* behavior under the "interacting with evidence" umbrella. For instance, evidential thresholds cannot. By themselves, they cannot explain differences in evidence-gathering, in alternative explanations generated, or in questions asked about the evidence. In fact, appealing to evidential thresholds does not yield a complete explanation even in the cases in which theorists appeal to them. For example, gritty people do not just require more evidence to

later.

change their minds on their chances of success. In addition to that, they often also robustly explain away counter-evidence to those beliefs; shape their trust policies in ways that allow them to devalue the testimony of people who do not believe in them; and focus their attention on signs of success. These aspects of behavior are not explained by evidential threshold settings.

To capture these facts, an answer to the Variation Question must meet the following desideratum:

> **Multi-Dimensionality Desideratum**: To account for variations in multiple dimensions of interacting with evidence and multiple epistemic parameters.

In other words, a full answer to the Variation Question requires accommodating variation in complex sets of epistemic parameters, and how they affect a broad range of behavior.

This might suggest appealing to differences in deep epistemic character—in epistemic virtues, vices, and global character traits (L. T. Zagzebski 1996) (e.g. open-mindedness, intellectual humility, arrogance)—to explain why people interact with evidence differently. Such a account seems promising for meeting the Multi-Dimensionality Desideratum because that global character traits are multi-track dispositions (Ryle 1949), i.e., they correspond to more than one pair of stimulus condition and manifestation. For this reason, they encompass settings in multiple epistemic parameters and are well-placed to explain a wide range of epistemic behavior.

However, such virtue-theoretic approaches face a substantive empirical challenge: the challenge from situationism (Harman 1999, Doris 2002, Alfano 2013, Fairweather and Alfano 2017). Having a global character trait requires robustly manifesting that trait across a wide range of conditions, not only in a narrow, hyper-specific range of conditions. For example, honesty requires reliably behaving in honest ways, not just behaving honestly when it's sunny and you've had a nice meal. But results in social psychology suggest that people do not robustly behave in trait-manifesting ways. Instead, normatively irrelevant situational influences—e.g. moderate social pressure, mood, framing—have substantial effects on behavior. As a consequence, global character traits (which, by definition, are robustly manifested across a wide range of conditions) are rare.

If the situationist is right, global character traits are not well-suited for addressing typical cases of variation in how people interact with evidence. More generally, situationism indicates that we need to leave space for the effect of contextual factors on epistemic behavior:

> **Context-Dependence Desideratum**: To accommodate the systematic dependence of our ways of interacting with evidence on context.

For instance, we need to leave space for the fact that we interact with evidence differently in different social contexts (e.g. in a philosophy seminar vs. at a bar with non-academic friends). We should also accommodate our tendency to reason in different ways when in a good mood and when feeling down (in exploratory vs. critical ways, respectively; Schaller and Cialdini 1990). Similarly, we should make space for the fact that taking up accuracy goals as opposed to wanting to defend one's cherished beliefs affects the ways in which we interact with evidence (Kunda 1990). At the same time, merely appealing to situational factors to explain epistemic behavior seems insufficient. In particular, different agents respond to situational factors in different ways, suggesting that we need to leave room for *the agent* in our explanations of epistemic behavior.

The discussion so far points to a gap in our theorizing. Existing theoretical tools—beliefs about the topic under discussion, performance mistakes, evidential threshold settings, deep character traits, situational factors—do not suffice to explain important instances of variation in interactions with evidence. In the rest of the paper, I will address this gap by introducing and developing an account of epistemic styles.

## 5.3 Epistemic Styles

### 5.3.1 Style: an overview

The notion of style has its primary home in aesthetics.[5] In one canonical use of the term, a style is a unified way of doing things: of dressing, gesturing, speaking, moving, and so on. Taking up a style is a matter of being disposed to do things in those ways. This is a descriptive notion of style, in that, on this notion of style, not all styles are (aesthetically) good. You can have a style without being stylish: you just need to have some unified way of doing things.[6]

Style shows up across a wide range of activities and domains: a flamboyant style can show up in flashy, glittery outfits, in pronounced facial expressions, and in throwing exuberant parties. At the same time, style is manifested in different ways across activities and domains. A flamboyant style will result in different outfits at a dance party and at a picnic, at least if one is sensitive to social norms.

When we talk of someone having a style, we mean that they do a number of things in the same way. As Arthur Danto notes, the notion of consistency at play here is not "formal" consistency:

> It is the consistency rather of the sort we invoke when we say that a rug does not fit with the other furnishings of the room, or a dish does not fit with the structure of a meal, or a man does not fit with his own crowd (Danto 1981, p. 207).

We cannot describe a style in a purely formal way, by listing abstract rules for combining different constituents (e.g. different items of clothing). Instead, the orthodox view is that what makes it the case that different actions are in the same style is that they are all done

[5] For classic discussions, see Sontag 1966, Danto 1981, Baxandall 1985, Robinson 1985, Wollheim 1987.

[6] As Riggle 2015 notes, there is also an evaluative notion of style on which style is an achievement. It is in this evaluative sense that some people are stylish and some people are not.

in ways that express (aspects of) the same psychological profile (Robinson 1985, Wollheim 1987). Behavior that is in a certain style shows or makes manifest (Green 2016) aspects of a unified psychological profile that the agent inhabits at the time. Actions such as dressing for a picnic or for a party, talking in a certain tone of voice, or characteristic gestures are in the same style in virtue of expressing the same psychological profile.

Note, however, that style is not fixed. People can and do shift styles, both over the course of their lives and across contexts. One's style does not express deep character, conceived as a set of stable, long-standing traits.[7] Having a style only requires having dispositions that are manifested in the contexts in which the person adopts the style, and inhabiting the corresponding psychological profile in those contexts. For example, someone who has a flamboyant style in their social but not professional life has the corresponding psychological features—a preference for the dramatic, a tendency for effusive displays of emotion, and a taste for boundary-pushing—in social contexts, but does not have them in professional contexts.[8]

To summarize this large literature, styles are ways of doing things, and taking up a style is a matter of having dispositions to do things in those ways. Styles are unified, with their unity deriving from the fact that they express aspects of the same psychological profile. Consequently, taking up a style involves contextually inhabiting that psychological profile.

Though the notion of style has its home in aesthetics, it has been put to use in explanatory projects in other domains. For example, linguists and philosophers of language have theorized at length about styles of linguistic expression and their social significance (Eckert 1989, Tannen et al. 2005), and feminist theorists have long encouraged us to attend to distinctive gendered "ways of knowing" (Belenky et al. 1986, Collins 2002, Gilligan 1993, Rooney 1991). Closer to my project here, philosophers of mind and action have provided

[7] Thanks to Elisabeth Camp and Thi Nguyen for illuminating discussion on this point.

[8] For more on how personality is affected by context, see Goffman 1978 on social roles, Rovane 2019 on ways of reasoning, C Thi Nguyen 2020b on modes of agency, and J. M. Morton 2014 on code-switching.

detailed accounts of style in intuitive cognition and action. Elisabeth Camp articulates *perspectives* as styles of intuitive thinking: packages of intuitive dispositions to notice, explain, and evaluate the world around us (Camp 2006, Camp 2013, Camp 2019, Camp 2020). And Thi C Thi Nguyen 2020b has developed the notion of *modes* (or styles) *of agency*, focused ways of being an epistemic agent that agents adopt in context-dependent ways.

These projects illustrate the explanatory power of the notion of style. For example, thinking about perspectives helps us understand the cognitive significance of linguistic devices such as slurs (Camp 2013) and metaphors (Camp 2006), the role of models in scientific inquiry (Camp 2020), and the structure of testimony (Fraser 2021). Modes of agency help explain our engagement with games and make-believe, how we shift values in the context of different activities, and the development of agency over time (C Thi Nguyen 2020b). Such explanations also raise new normative questions about which perspectives and modes of agency we ought to adopt. Similarly, I will show that epistemic styles help us better understand a wide range of epistemic behavior and raise new normative questions about which epistemic styles we ought to adopt.

### 5.3.2 Style in an epistemic key

Applying the points in the last sub-section to the epistemic domain, here is my definition of epistemic style:

> **Epistemic Style**: An epistemic style is a way of interacting with evidence that expresses (aspects of) a unified set of epistemic parameters.

Taking up an epistemic style is a matter of having the dispositions that constitute that epistemic style and setting epistemic parameters accordingly. Epistemic styles are flexible: people can and do shift their style over time and across contexts, by re-setting their epistemic parameters and adopting the corresponding epistemic dispositions.

It will be easier to get a grip on the notion of epistemic style by considering some examples.

Consider, first, the paranoid style, introduced by Richard Hofstadter as a style for "angry minds," expressive of "heated exaggeration, suspiciousness, and conspiratorial fantasy" (Hofstadter 2012). Interacting with evidence in the paranoid style is a matter of doing so in ways that express that psychological/epistemic profile. In recent work, Rachel Fraser 2020 has further articulated the epistemic parameters expressed in the paranoid style as involving "a coupling of Cartesian paranoia ["refusal to allow that the evidence really guarantees what it appears to show"] with a very unCartesian passional structure: epistemic fear of missing out, or FOMO" (Fraser 2020), characterized by extreme epistemic risk-seeking.[9]

A second instructive example is the rationalist style, a way of interacting with evidence that the self-proclaimed rationalist community strives to inculcate and promote. This epistemic style is characterized by adhesion to Bayesian reasoning and by a "scout mindset" (Galef 2021), rooted in curiosity and willingness to change one's mind. It also encompasses a tendency to contrarianism and openness to exploring views regardless of the moral costs of doing so. According to critics, it also involves intellectual arrogance, manifested in deep trust of one's judgments and unwillingness to defer to others (Metz 2021). Rationalism is not just a set of epistemic commitments: it is meant to be inhabited and made manifest in how agents actually interact with evidence. In other words, the ideal rationalist not only endorses the epistemic commitments of the movement, but also adopts the corresponding epistemic style, setting their epistemic parameters so as to be disposed to reason according to Bayes' theorem, seeking out alternative explanations in a fairly unconstrained way, omnivorously consuming data and statistics, responding to personal testimony with an attitude of skepticism, and so on.

For a third example, consider the epistemic practices discussed by Patricia Hill Collins 2002 as characteristic of Black feminists in the United States. Collins argues that Black feminists typically take "lived experience as a criterion for credibility" (Collins 2002, p. 258),

[9] I am not committed to this analysis capturing all cases of the paranoid style. I am just employing this analysis to articulate an example of an epistemic style. The same caveat applies to the two examples below.

preferring testimony from people with relevant experiences of oppression over impersonal descriptions and testimony that is conveyed with emotion over coldly expressed points. They are disposed to seek out and value a wide range of distinctive perspectives. And they place a high value on dialogue, as opposed to more combative ways of interacting, with "new knowledge claims...usually developed through dialogues with other members of the community" (Collins 2002, p. 260). To put it differently, taking up the Black feminist epistemic style involves having dispositions to omnivorously seek out personal narratives from a range of different social positions, to take seriously evidence provided in an emotionally invested way, and to change one's mind through dialogue.

These are just three examples of epistemic styles. There are many more. Any unified way of interacting with evidence where the unity derives from the expression of epistemic parameter settings is an epistemic style.

Epistemic styles are commonplace. In part, this is because having an epistemic style does not require having reflective epistemic commitments, and one's style can come apart from whatever reflective epistemic commitments one has. Epistemic style does not express epistemic commitments: it expresses epistemic parameter settings. These parameters include many features that epistemologists have discussed at length. They include: Jamesian preferences for collecting true beliefs versus avoiding false beliefs (James 1979); risk preferences with respect to epistemic goods (Buchak 2013); and weightings of theoretical values (e.g. observational adequacy vs. fit with common sense; Douven 2009, T. Kelly 2013, Willard-Kyle 2017). They also include evidential policies, which collect the agent's evidential thresholds for a range of beliefs (Paul and J. Morton 2018), and trust policies, which set how one allocates epistemic trust in other agents.

I do not mean this list of epistemic parameters to be exhaustive or definitive. The point is that the kinds of aspects of psychology expressed in epistemic style are familiar from epistemology. At the same time, though epistemologists have discussed all of these parameters, they have failed to notice that they cluster into epistemic styles—such as the paranoid, rationalist, and Black feminist styles.[10] The epistemic behavior of agents who have these styles is not well-understood in terms of isolated epistemic parameters. To get a holistic sense of their behavior, we need to appeal to epistemic styles. Specifically, as I will now show, appealing to epistemic styles to explain ways of interacting with evidence meets all desiderata outlined in section 5.2.

## 5.4 Answering the Variation Question

I set out to explain systematic differences in ways of interacting with evidence. I argued in section 5.2 that existing explanatory tools do not suffice. I will now argue that epistemic styles do the job. Specifically, I will argue that: (a) when an individual systematically interacts with evidence in different ways when placed in different contexts, this is typically due to a shift in epistemic style; and, (b) when two people interact with evidence in systematically different ways, a difference in epistemic styles is typically part of the explanation.

This account leaves space for isolated epistemic parameters and global character traits to play an explanatory role. Specifically, there are cases in which isolated epistemic parameters explain the way in which an agent interacts with evidence. But, given the kinds of inter-connected patterns of variation we encounter, these will be marginal cases. Similarly, global epistemic character traits explain some patterns of behavior. But, given the sensitivity of our epistemic behavior to situational factors, such traits will be explanatory only in unusual cases.

I will now show that appealing to epistemic styles meets all desiderata I outlined in section 5.2.

[10] I leave open what the source of this clustering is. In some cases, it is internal: plausibly, the epistemic parameter settings of the paranoid style fit together because they all derive from a single psychological trait, e.g. suspiciousness. In other cases, it may be external, coming from social groups coordinating around packages of ways of interacting with evidence, as seems to be the case for the rationalist style.

### 5.4.1 Meeting the desiderata

I will start with the Epistemic Parameters, Multi-Dimensionality, and Context-Dependence Desiderata. In doing so, I will argue that appealing to epistemic styles is well-supported by empirical findings about how we interact with evidence. I will then discuss how the view meets the Predictive Validity and Intelligibility Desiderata, which will illuminate the practical and social benefits of the account.

> **Epistemic Parameters Desideratum**: To account for the fact that at least some differences in interactions with evidence are the result of differences in epistemic parameter settings.

Epistemic styles express settings in (unified sets of) epistemic parameters. When differences in interactions with evidence are the result of differences in epistemic style, they are also describable as the result of differences in epistemic parameter settings. For this reason, appealing to epistemic styles satisfies the Epistemic Parameters Desideratum.

> **Multi-Dimensionality Desideratum**: To account for variations in all dimensions of interacting with evidence, and in multiple epistemic parameters.

Epistemic styles are ways of interacting with evidence, where interacting with evidence covers changing one's beliefs in the light of evidence, gathering evidence, asking questions, considering alternative explanations, and so on. For this reason, appealing to epistemic styles can cover variations in all dimensions of interacting with evidence. Further, epistemic styles express sets of epistemic parameters, not a single parameter. Consequently, appealing to epistemic styles captures variations in multiple epistemic parameters.

> **Context-Dependence Desideratum**: To accommodate the systematic dependence of our ways of interacting with evidence on context.

Unlike the global epistemic character traits of virtue epistemological approaches, an agent's epistemic style can change across contexts. For this reason, appealing to epistemic styles can accommodate the context-dependence of our ways of interacting with evidence. As such, it can take into account situationist results (discussed in section 5.2).

Situationism, however, is primarily a negative view. In contrast, I provide a systematic framework in which to think of the role of situational factors, one which leaves space for our cognitive agency. In this framework, the influence of situational factors does not make agents empty vehicles through which context operates. Instead, such factors lead agents to (epistemically) code-switch (J. M. Morton 2014). Contextual factors shape and constrain which epistemic style agents take up.[11] And epistemic style expresses *the agent's* epistemic parameter settings in that context. This leaves space for agency in shaping interactions with evidence.[12]

The overall picture of the role of context is a *fragmentationist* one. Fragmentationists about belief hold that different sets of beliefs (fragments) guide action in different contexts, and appeal to this to explain cases of inter-context behavioral inconsistency.[13] Similarly, different epistemic styles are activated in different contexts. This explains differences in epistemic behavior across contexts.

For example, which epistemic style we take up is in general sensitive to that of others around us. As Collins 2002 argues, many of the features of the Black feminist epistemic style are derived from practices of interaction characteristic of Black communities in the United States. Similarly, people often develop the paranoid style as the result of immersion in conspiracist communities. And the rationalist style is actively promoted and taught by the rationalist community, through textbooks, workshops, and online community spaces.[14]

[11] Such factors can also cause performance mistakes or, more generally, lead agents to act out of style.

[12] Developing an account of the agency we have over our epistemic styles is beyond the scope of this paper.

[13] See D. K. Lewis 1982, Egan 2008b and the essays in Borgoni et al. 2021 for more on belief fragmentation.

[14] The Center for Applied Rationality (https://www.rationality.org/) offers workshops, at $4,900 for four and a half days. And there are a range of online manuals to this style, including Eliezer Yukowski's *Harry*

For a second example of how appeal to epistemic style captures the effects of situational factors, consider the effects of mood. A positive mood tends to make us more exploratory, and a negative mood more critical and detail-oriented (Schaller and Cialdini 1990). This phenomenon is well-captured in terms of a shift in epistemic style. Moods elicit distinctive epistemic styles: they lead us to re-set epistemic parameters and to adopt corresponding dispositions. Similar remarks apply to the way in which accuracy vs. defensiveness goals constrain how we interact with evidence (Kunda 1990).

In general, non-epistemic factors affect how we interact with evidence indirectly, by triggering shifts in epistemic style. This leaves room for empirical investigation the elicitation conditions for epistemic styles (i.e. the conditions in which a specific epistemic style is elicited in an agent). And it suggests that fragmentationists about belief need an additional variable in their theory: not just which beliefs are active, but also which epistemic style is at play.

The fact that appealing to epistemic styles meets the Epistemic Parameters, Multi-Dimensionality, and Context-Dependence desiderata shows that appealing to epistemic styles is well-suited to describe our interactions with evidence. This makes it unsurprising that appeal to epistemic styles helps us predict how agents will interact with evidence:

> **Predictive Validity Desideratum**: To put us in a position to predict how others will interact with a range of evidence, if we have relevant information.

Knowing someone's style, in general, helps us predict their (style-related) behavior. It helps us predict how they will dress, speak, react to others, and so on, in contexts in which they have that style. Similarly, knowing someone's epistemic style helps us predict how they will interact with evidence. Knowing someone's epistemic style involves knowing how they are disposed to interact with evidence in a context: which evidence they are

*Potter and the Methods of Rationality*, and online spaces committed to this style of reasoning, such as the online forum Less Wrong (https://www.lesswrong.com/), "a community blog devoted to refining the art of rationality."

disposed to take seriously, which evidence is likely to change their minds, which circumstances are likely to elicit evidence-gathering behavior, and so on. If we know someone's epistemic style in a context and their relevant beliefs, we are well-equipped to predict how they will respond to evidence in that context. For this reason, the Epistemic Styles View meets the Predictive Validity Desideratum.

Such predictions are not infallible. Styles are usually compatible with multiple responses to evidence. And dispositions are not always manifested in their elicitation conditions. They can be masked (Johnston 1992, Bird 1998): much like a fragile glass might fail to break when struck because it is carefully wrapped, a person with a paranoid epistemic style might fail to come up with a conspiratorial explanation for the evidence because they are too tired.

The fact that the view meets the Predictive Validity Desideratum is practically significant. It provides tools for (at least partially) addressing the difficulties in rational persuasion and joint deliberation I mentioned in section 5.2. You can canvass your knowledge of epistemic style to select evidence that your interlocutor will find persuasive, to determine how much evidence to offer, and to anticipate and pre-empt objections to your arguments that they are likely to bring up.

For example, if you are trying to persuade someone who interacts with evidence in the paranoid style, you should expect them to find conspiratorial explanations highly salient, to strongly prefer evidence from more informal sources than from the mainstream media, and to set high evidential thresholds for changing their mind across the board. Armed with this knowledge, you can select more persuasive evidence to offer, anticipate alternative explanations for that evidence, and persevere in a way that is sensitive to their high evidential thresholds. Alternatively, you might decide to pass on seriously engaging until you are in a context where they will take up a more receptive epistemic style.[15]

Knowledge of epistemic style makes a distinctive contribution here. If your interlocu-

[15] Thanks to Christopher Willard-Kyle for helpful discussion.

tor adopts the rationalist style, you would be practically well-served by offering evidence from academic sources, expecting them to reason carefully through probabilistic evidence and to be open to alternative explanations that may have morally dubious implications, and so on.[16]

This leaves one final desideratum to address:

> **Understanding Desideratum**: To put us in a position to understand interactions with evidence.

Understanding comes in different kinds and degrees (Grimm 2016). My discussion below cannot do justice to all varieties of understanding. That said, I will try to make the case that knowledge of epistemic style contributes to important kinds of understanding.

One kind of understanding—the kind of understanding characteristic of the natural sciences—consists (roughly) in intellectually grasping a causal model of the factors which lie behind the target's behavior, in a way that allows for good predictions (Kitcher 1989).

Knowing an agent's epistemic style in a context is a matter of knowing how they are disposed to interact with evidence. In itself, knowledge of dispositions does not yield knowledge of the cognitive basis of these dispositions. At the same time, epistemic styles express epistemic parameter settings. If one also comes to have a sense for these parameter settings, one comes to this kind of understanding of how an agent interacts with evidence. There is good reason to think that knowing someone's style puts one in a position to determine what the background epistemic parameter settings are. As McGeer 2007b puts it, we are "inveterate mentalizers" (McGeer 2007b, p. 137): we find it natural to understand all sorts of behavior in psychological terms. Our knowledge of an agent's style tends to be accompanied by a sense of the psychological features that those surface dispositions

[16] Of course, epistemic style is only one factor among many that contribute to successful persuasion. For example, effective engagement will require attention to how evidence is presented (Tversky and Kahneman 1981), not only to which evidence one presents. Further, I leave open important ethical questions: how do we employ knowledge of epistemic style in respectful, non-manipulative ways? What are moral constraints on using our knowledge of others' epistemic style? My discussion here serves as a prolegomenon to such questions. To address them, we first need to bring epistemic styles into clear focus.

express. For this reason, knowledge of style typically puts us in a position to achieve naturalistic understanding of others.

We often want to understand others in ways that go beyond such naturalist understanding. We want to understand others *as agents* who act for reasons, whose behavior is rationally intelligible. As Grimm 2016 notes, such understanding requires seeing others not (merely) as causal mechanisms. Such understanding is holistic. As Iris Murdoch put it, when we understand other people,

> we do not consider only their solutions to specifiable practical problems, we consider something elusive which may be called their total vision of life... in short, the configurations of their thought which show continually in their reactions and conversation (Murdoch 1956, p. 39).

Knowing others' epistemic styles can make a distinctive contribution to understanding them in this rich humanistic sense. It involves having a sense of "the configurations of their thought" which show in their interactions with evidence: in this case, of the epistemic parameter settings that are expressed in how they interact with evidence. In virtue of knowing someone's epistemic style, we come to understand them *qua* epistemic agents who (in that context, on that topic) live by certain epistemic values.[17]

This is a significant result. Grasp of epistemic style can rescue us from seeing others as profoundly irrational, specifically, as agents whose epistemic behavior is purely determined by irrelevant factors ( such as strong emotions, partisan affiliations, or vicious motivation). By appeal to epistemic styles, we can acknowledge that such factors can and do play a role in how people interact with evidence: but they do so by reshaping their epistemic parameter settings. Crucially, we can make genuine sense of their epistemic behavior in light of such settings.

[17] They may have long-standing values which are not manifested in the style they take up. Knowledge of style will not help us make sense of their behavior by reference to *those* values. That is the right result when those values are not in fact expressed in behavior.

Note that, on this view, one can make genuine sense of others' epistemic behavior without viewing them as epistemically rational. This goes against views that postulate a constitutive connection between folk-psychological intelligibility and epistemic rationality (D. Davidson 1973). On my view, we can see how a pattern of behavior makes sense in the light of a set of epistemic parameters while thinking that setting one's epistemic parameters in those ways (in some, perhaps actual, contexts) leads to irrational interactions with evidence. In other words, knowledge of epistemic style enables us to make rational sense of epistemic behavior in light of specific parameter settings, but this need not involve claiming that such behavior is rational.

The kind of understanding of others that I have discussed so far is distanced and third-personal. It involves making sense of others' responses to evidence intellectually, by seeing how they express an epistemic profile. There are, however, more involved or empathetic kinds of understanding of others that one may want to attain. We may want to get others' parameter settings "from the inside" by simulating them (Goldman 2006, Maibom 2007), or to be able to see those parameter settings as good or choiceworthy (Grimm 2016).

Such kinds of understanding appear especially socially and politically valuable, because they reduce disdain for others and help resolve deep conflicts (Hannon 2020). These benefits are particularly significant in political deliberation, which requires tolerance, mutual respect, and openness towards others. Empathetic understanding can function as an antidote to the kind of polarization (Benkler et al. 2018) that often dominates political contexts, and thereby enable people to have better conversations and reap the benefits of collective deliberation.

Merely knowing someone's epistemic style does not suffice for empathetic understanding: achieving empathetic understanding requires additional imaginative and perspectival work. Nevertheless, knowing someone's epistemic style is an important ingredient for this work. We need to have a sense of how others have set their epistemic parameters if we are to simulate them or come to see them as choiceworthy from some perspective.

All things considered, appealing to epistemic styles meets the Understanding Desideratum. Knowing someone's epistemic style puts us in a position to begin to understand the causal structure behind their interactions with evidence; it helps us make sense of agents' interactions with evidence at a personal level; and it provides us with knowledge which we can canvass to arrive at empathetic understanding.[18]

## 5.5 Upshots of Epistemic Styles for Epistemology

I have argued that epistemic styles make a crucial contribution to how real-world agents interact with evidence. Insofar as epistemologists are interested in providing tools for assessing real-world epistemic agents, they have good reason to be interested in epistemic styles.

In criticizing virtue epistemology, situationists press a similar point about the importance of attending to how agents actually interact with evidence (Fairweather and Alfano 2017). I go beyond the situationist critique in providing a new object for epistemic assessment: epistemic styles. I have argued that individuals take up epistemic styles, and that such styles are behind our interactions with evidence. If this is right, then we can begin to build a more applicable theory of epistemic assessment that improves on virtue epistemology by focusing on assessing epistemic styles instead of global character traits.[19] At a practical level, we should re-allocate our attention from thinking about how to promote epistemic virtues to thinking about how to inculcate good epistemic styles.[20]

Similarly, my discussion suggests that we need to expand our discussions of moral and pragmatic encroachment.[21] Such discussions tend to focus on how individuals ought

[18] Much like the discussion of rational engagement above, this discussion of the understanding-related benefits of sensitivity to epistemic style is only preliminary. When and to what extent sensitivity to epistemic style has these benefits is a difficult empirical question.

[19] See Lasonen-Aarnio 2020 for an account of how to epistemically assess dispositions that is relevant here, given that styles are packages of dispositions.

[20] Thanks to Miranda Fricker for suggesting this point.

[21] Thanks to Quill Kukla for suggesting this point.

to shift isolated epistemic parameters—their evidential thresholds or spheres of relevant alternatives—in the light of moral or practical factors.[22] But moral and pragmatic factors do not only affect how individuals set evidential thresholds or spheres of relevant alternatives. They lead individuals to *shift epistemic styles*, adjusting a wide range of epistemic parameters and behavior. To provide norms on encroachment, we need to assess shifts in epistemic style, not only in isolated epistemic parameters.

Further, a theory of the epistemic assessment of epistemic styles may help us assess epistemic conduct in cases of deep disagreement.[23] An intriguing hypothesis is that such disagreement is (at least in some cases) sustained by differences in epistemic style. Where that is the case, assessing agents' conduct will involve assessing their epistemic styles and addressing questions about when one ought to change one's epistemic style.

At the level of communal disagreement, epistemic styles might help us theorize about epistemic bubbles and echo chambers (informational structures that omit or actively exclude relevant information, respectively; see C. Thi Nguyen 2020a). Perhaps some such informational structures are partly sustained by divergences in epistemic style at a community-level. If that is right, then dissolving these structures might require community-level shifts in epistemic style. Theorists interested in these phenomena should think through when shifts in epistemic style are appropriate and what are good means to bring them about.

Finally, appeal to epistemic styles may also help us understand when and why different ways of knowing (Belenky et al. 1986, Collins 2002, Gilligan 1993, Rooney 1991) are valuable. Their value might in part be explainable in terms of the value of different epistemic styles. One important benefit of approaching this question through the lenses of epistemic style is that doing so avoids essentialism about modes of epistemic engagement, that is, it avoids seeing such modes of engagement as innate or essential to members of certain social groups. Epistemic styles are packages of dispositions that one can take

[22] See Bolinger 2020 for an overview of different kinds of encroachment proposed in the literature. Detailed discussion of all versions of the encroachment view is beyond the scope of this paper.

[23] See Frances and Matheson 2019 for an overview.

up and abandon, not innate or immutable traits of individuals. For this reason, the claim that marginalized social groups have characteristic epistemic styles is non-essentializing, leaving space to recognize the role of social factors in the construction and adoption of epistemic styles.

### 5.6 Conclusion

Why do we find diversity in how people interact with evidence? To address this question, I introduced and developed the notion of epistemic styles: unified ways of interacting with evidence that express (settings of) epistemic parameters which agents can flexibly take up. I argued that appealing to differences in epistemic style best accounts for cases of systematic variation in interactions with evidence.

Though I introduced the notion of epistemic style to address a descriptive question—what explains people's distinctive ways of interacting with evidence—the notion can be put to work to reshape our normative theorizing. It can help us think through important questions in epistemology—for example, about disagreement, cognitive diversity, and echo chambers. More generally, epistemic styles provide a framework within which to theorize about epistemic assessment.

## ACKNOWLEDGMENT OF PREVIOUS PUBLICATIONS

**P1** Flores, Carolina (2021). Delusional evidence-responsiveness. *Synthese* 199 (3-4):6299-6330..

**P2** Flores, Carolina (forthcoming). Epistemic styles. *Philosophical Topics.*

# REFERENCES

Achen, C. H., & Bartels, L. M. (2016). *Democracy for realists*. Princeton University Press.

Alcoff, L. M. (2007). Epistemologies of ignorance: Three types. In S. S. N. Tuana (Ed.), *Race and epistemologies of ignorance*. SUNY Press.

Alexander, M. [Michelle]. (2010). *The new jim crow: Mass incarceration in the age of colorblindness*. The New Press.

Alexander, M. [M.P.], Stuss, D., & Benson, D. (1979). Capgras' syndrome: A reduplicative phenomenon. *Neurology*, *29*, 334–339.

Alfano, M. (2013). *Character as moral fiction*. Cambridge University Press.

Allport, G. (1954). *The nature of prejudice*. Addison-Wesley.

Amenta, E., & Polletta, F. (2019). The cultural impacts of social movements. *Annual Review of Sociology*, *45*, 279–299.

Anderson, C., Lepper, M., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology*, *39*, 1037–1049.

Anderson, E. (2010). *The imperative of integration*. Princeton University Press.

Anderson, E. (2012). Epistemic justice as a virtue of social institutions. *Social epistemology*, *26*(2), 163–173.

Anderson, E. (2014). Social movements, experiments in living, and moral progress: Case studies from britain's abolition of slavery.

Antony, L. (2016). Bias: Friend or foe? Reflections on Saulish skepticism. *Implicit bias and philosophy*, *1*, 157–190.

Appelbaum, P. S., Robbins, P. C., & Roth, L. H. (1999). Dimensional approach to delusions: Comparison across types and diagnoses. *American Journal of Psychiatry*, *156*(12), 1938–1943.

Arendt, H. (1989). *Lectures on Kant's political philosophy*. University of Chicago Press.

Ashwell, L. (2010). Superficial dispositionalism. *Australasian Journal of Philosophy*, *88*(4), 635–653.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders (dsm-5®)*. American Psychiatric Pub.

Avery, B., & Lu, H. (2021). Ban the box: U.S. cities, counties, and states adopt fair hiring policies.

Ayala-López, S., & Beeghly, E. (2020). Explaining injustice: Structural analysis, bias, and individuals. *An introduction to implicit bias* (pp. 211–232). Routledge.

Baxandall, M. (1985). *Patterns of intention: On the historical explanation of pictures*. Yale University Press.

Bayne, T., & Fernández, J. (2010). *Delusion and self-deception: Affective and motivational influences on belief formation*. Psychology Press.

Bayne, T., & Pacherie, E. (2004). Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry, & Psychology*, *11*(1), 1–11.

Bayne, T. J., & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind and Language*, *20*(2), 163–88.

Begby, E. (2021a). Evidential preemption. *Philosophy and Phenomenological Research*.

Begby, E. (2021b). *Prejudice: A study in non-ideal epistemology*. Oxford University Press.

Belenky, M. F., Clinchy, B. M., Goldberger, N. R., Tarule, J. M. et al. (1986). *Women's ways of knowing: The development of self, voice, and mind* (Vol. 15). Basic books New York.

Bell, V., Halligan, P. W., & Ellis, H. D. (2008). Are anomalous perceptual experiences necessary for delusions? *The Journal of nervous and mental disease*, *196*(1), 3–8.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in american politics*. Oxford University Press.

Bentall, R. P., Corcoran, R., Howard, R., Blackwood, N., & Kinderman, P. (2001). Persecutory delusions: A review and theoretical integration. *Clinical psychology review*, *21*(8), 1143–1192.

Berrios, G. (1991). Delusions as 'wrong beliefs': A conceptual history. *British Journal of Psychiatry*, *159*, 6–13.

Bicchieri, C. (2016). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.

Bird, A. (1998). Dispositions and antidotes. *The Philosophical Quarterly*, *48*(191), 227–234.

Bird, A. (2007). *Nature's metaphysics: Laws and properties.* Oxford University Press on Demand.

Bishop, B. (2009). *The big sort: Why the clustering of like-minded america is tearing us apart.* Houghton Mifflin Harcourt.

Bisiach, E. (1988). Language without thought. In L. Weiskrantz (Ed.), *Thought without language* (pp. 464–84). Oxford University Press.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Block, N., & Fodor, J. A. [Jerry A.]. (1972). What psychological states are not. *Philosophical Review*, *81*(April), 159–81.

Blount, G. (1986). Dangerousness of patients with Capgras syndrome. *Nebraska Medical Journal*, *71*(207).

Bolinger, R. J. (2020). Varieties of moral encroachment. *Philosophical Perspectives*, *34*(1), 5–26.

Borgoni, C., Kindermann, D., & Onofri, A. (2021). *The fragmented mind.* Oxford University Press.

Bortolotti, L. (2005a). Delusions and the background of rationality. *Mind and Language*, *20*(2), 189–208.

Bortolotti, L. (2005b). Intentionality without rationality. *Proceedings of the Aristotelian Society*, *105*(3), 385–392.

Bortolotti, L. (2009). *Delusions and other irrational beliefs.* Oxford University Press.

Bortolotti, L. (2018). Delusion. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2018). Metaphysics Research Lab, Stanford University.

Bortolotti, L., & Miyazono, K. (2015). Recent work on the nature and development of delusions. *Philosophy Compass*, *10*(9), 636–645.

Bovet, P., & Parnas, J. (1993). Schizophrenic delusions: A phenomenological approach. *Schizophrenia bulletin*, *19*(3), 579–597.

Brandenburg, D. (2018). The nurturing stance: Making sense of responsibility without blame. *Pacific Philosophical Quarterly*, *99*, 5–22.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment.* Harvard university press.

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

Breinlinger, S., & Kelly, C. (2014). *The social psychology of collective action.* Taylor & Francis.

Brennan, J. (2016). *Against democracy.* Princeton University Press.

Broome, M., Johns, L., Valli, I., Woolley, J., Tabraham, P., Brett, C., Valmaggia, L., Peters, E., Garety, P., & McGuire, P. (2007). Delusion formation and reasoning biases in those at clinical high risk for psychosis. *The British Journal of Psychiatry*, *191*(S51), s38–s42.

Brown, R. (2000). Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology*, *30*(6), 745–778.

Brownstein, M., Kelly, D., & Madva, A. (2021). Individualism, structuralism, and climate change. *Environmental Communication*, 1–20.

Bruce, V., & Young, A. [Andy]. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Buchak, L. (2013). *Risk and rationality.* Oxford University Press.

Burge, T. (2010). *Origins of objectivity.* Oxford University Press.

Burgess, A., Cappelen, H., & Plunkett, D. (2019). *Conceptual engineering and conceptual ethics.* Oxford University Press, USA.

Burgess, A., & Plunkett, D. (2013). Conceptual ethics i. *Philosophy Compass*, *8*(12), 1091–1101.

Callahan, L. F. (2021). Epistemic existentialism. *Episteme*, 1–16.

Camp, E. (2006). Metaphor and that certain 'Je ne sais quoi'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *129*(1), 1–25.

Camp, E. (2013). Slurring perspectives. *Analytic Philosophy*, *54*(3), 330–349.

Camp, E. (2015). Logical concepts and associative characterizations. In E. Margolis & S. Laurence (Eds.), *Conceptual mind.* MIT Press Cambridge, MA.

Camp, E. (2017). Perspectives in imaginative engagement with fiction. *Philosophical Perspectives*, *31*(1), 73–102.

Camp, E. (2019). Perspectives and frames in pursuit of ultimate understanding. *Varieties of Understanding*, 17–46.

Camp, E. (2020). Imaginative frames for scientific inquiry: Metaphors, telling facts, and just-so stories. *The scientific imagination* (pp. 304–336). Oxford University Press.

Camp, E. (2022). Agency, stability, and permeability in "games". *Journal of Ethics and Social Philosophy*.

Camp, E., & Flores, C. (2022). "That's all you really are": Centering social identities without essentialist beliefs. *Mind, language, and social hierarchy: Constructing a shared social world*. Oxford University Press.

Campbell, J. (2001). Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology*, *8*(2), 89–100.

Capgras, J., & Reboul-Lachaux, J. (1994). L'illusion des' sosies' dans un délire systématisé chronique. *History of Psychiatry*, *5*(17), 119–133.

Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1–18.

Chapman, R. K. (2002). First person account: Eliminating delusions. *Schizophrenia Bulletin*, *28*(3), 545–553.

Chatterjee, A. (1996). Anosognosia for hemiplegia: Patient retrospections. *Cognitive Neuropsychiatry*, *1*(3), 221–237.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of educational research*, *63*(1), 1–49.

Choi, S. (2005). Do categorical ascriptions entail counterfactual conditionals? *The Philosophical Quarterly*, *55*(220), 495–503.

Choi, S., & Fara, M. (2018). Dispositions. *The Stanford Encyclopedia of Philosophy*.

Churchland, P. M. (1981). Eliminative materialism and propositional attitudes. *the Journal of Philosophy*, *78*(2), 67–90.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Cohen, D., & Handfield, T. (2007). Finking Frankfurt. *Philosophical Studies*, *135*(3), 363–374.

Cohen, G. L., Aronson, J., & Steele, C. M. (2000). When beliefs yield to evidence: Reducing biased evaluation by affirming the self. *Personality and social psychology bulletin*, *26*(9), 1151–1164.

Cohen, G. L., Sherman, D. K., Bastardi, A., Hsu, L., McGoey, M., & Ross, L. (2007). Bridging the partisan divide: Self-affirmation reduces ideological closed-mindedness and inflexibility in negotiation. *Journal of personality and social psychology*, *93*(3), 415.

Collins, P. H. (2002). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

Coltheart, M. [Max]. (2005). Conscious experience and delusional belief. *Philosophy, Psychiatry, & Psychology*, *12*(2), 153–157.

Coltheart, M. [Max]. (2007). Cognitive neuropsychiatry and delusional belief. *Quarterly journal of experimental psychology (2006)*, *60*(8), 1041–1062.

Coltheart, M. [Max], Langdon, R., & McKay, R. (2011). Delusional belief. *Annual review of psychology*, *62*, 271–298.

Cooper, J. (2007). *Cognitive dissonance: 50 years of a classic theory*. Sage.

Corlett, P. R., Krystal, J. H., Taylor, J. R., & Fletcher, P. C. (2009). Why do delusions persist? *Frontiers in human neuroscience*, *3*, 12.

Cotard, J. (1880). Du délire hypochondriaque dans une forme grave de la méleancolie anxieuse. *Annales medico-psychologiques*, *4*, 168–174.

Craig, E. (1991). *Knowledge and the state of nature: An essay in conceptual synthesis*. Clarendon Press.

Currie, G., & Jureidini, J. (2001). Delusion, rationality, empathy: Commentary on martin davies et al. *Philosophy, Psychiatry, & Psychology*, *8*, 159–162.

Currie, G., & Ravenscroft, I. (2002). *Recreative minds: Imagination in philosophy and psychology*. Oxford University Press.

Danto, A. C. (1981). *The transfiguration of the commonplace: A philosophy of art*. Harvard University Press.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in experimental social psychology*, *47*, 233–279.

Davidson, D. (1973). Radical interpretation. *Dialectica*, 313–328.

Davidson, D. (1982). Rational animals. *dialectica*, *36*(4), 317–327.

Davidson, D. (1985). Incoherence and irrationality. *Dialectica*, *39*(4), 345–54.

Davidson, L. J., & Kelly, D. (2020). Minding the gap: Bias, soft structures, and the double life of social norms. *Journal of Applied Philosophy*, *37*(2), 190–210.

Davies, A. M. A., & Davies, M. (2009). Explaining pathologies of belief. In M. Broome & L. Bortolotti (Eds.), *Psychiatry as cognitive neuroscience: Philosophical perspectives.* Oxford University Press.

Davies, M., & Coltheart, M. [Max]. (2000). Pathologies of belief. In M. Davies & M. Coltheart (Eds.), *Pathologies of belief.* Wiley-Blackwell.

Dawson, E., Savitsky, K., & Dunning, D. (2006). “don’t tell me, i don’t want to know”: Understanding people’s reluctance to obtain medical diagnostic information. *Journal of Applied Social Psychology*, *36*, 751–768.

Dellaposta, D. (2020). Pluralistic collapse: The “oil spill” model of mass opinion polarization. *American Sociological Review*, *85*(3), 507–536.

Dennett, D. C. (1981). True believers: The intentional strategy and why it works. In A. F. Heath (Ed.), *Scientific explanation: Papers based on herbert spencer lectures given in the university of oxford* (pp. 150–167). Clarendon Press.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology*, *75*(1), 53.

Dixon, J., Levine, M., Reicher, S., & Durrheim, K. (2012). Beyond prejudice: Are negative evaluations the problem and is getting us to like one another more the solution? *Behavioral and Brain Sciences*, *35*(6), 411–425.

Dogramaci, S., & Horowitz, S. (2016). An argument for uniqueness about evidential support. *Philosophical Issues*, *26*(1), 130–147.

Doleac, J. L., & Hansen, B. (2016). *Does “Ban the Box” help or hurt low-skilled workers? statistical discrimination and employment outcomes when criminal histories are hidden* (tech. rep.). National Bureau of Economic Research.

Doris, J. M. (2002). *Lack of character: Personality and moral behavior.* Cambridge University Press.

Dorst, K. (2019). Why rational people polarize. *The Phenomenal World*.

Douven, I. (2009). Uniqueness revisited. *American Philosophical Quarterly*, *46*(4), 347–361.

Dover, T. L., Major, B., & Kaiser, C. R. (2014). Diversity initiatives, status, and system-justifying beliefs: When and how diversity efforts de-legitimize discrimination claims. *Group Processes & Intergroup Relations*, *17*(4), 485–493.

Dovidio, J. F., Isen, A. M., Guerra, P., Gaertner, S. L., & Rust, M. (1998). Positive affect, cognition, and the reduction of intergroup bias. *Intergroup cognition and intergroup behavior* (pp. 337–366). Lawrence Erlbaum Associates Publishers.

Dozois, D., & Dobson, K. (Eds.). (2010). *Handbook of cognitive behavioral therapies*. Guilford Press.

Dryzek, J. S. (2002). *Deliberative democracy and beyond: Liberals, critics, contestations*. Oxford University Press on Demand.

Dub, R. (2017). Delusions, acceptances, and cognitive feelings. *Philosophy and Phenomenological Research*, *94*(1), 27–60.

Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Egan, A. (2008a). Imagination, delusion, and self-deception. In T. Bayne & J. Fernandez (Eds.), *Delusion and self-deception: Affective and motivational influences on belief formation (macquarie monographs in cognitive science)*. Psychology Press.

Egan, A. (2008b). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47–63.

Elkin, R. A., & Leippe, M. R. (1986). Physiological arousal, dissonance, and attitude change: Evidence for a dissonance-arousal link and a “don’t remind me” effect. *Journal of personality and social psychology*, *51*(1), 55.

Elliot, A., & Devine, P. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, *67*(3), 384–394.

Ellis, H. D., & Lewis, M. B. (2001). Capgras delusion: A window on face recognition. *Trends in cognitive sciences*, *5*(4), 149–156.

Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *The British Journal of Psychiatry*, *157*(2), 239–248.

Estlund, D. (2009). *Democratic authority: A philosophical framework.* Princeton University Press.

Fairweather, A., & Alfano, M. (2017). *Epistemic situationism.* Oxford University Press.

Fairweather, A., Zagzebski, L. T., Zagzebski, L. et al. (2001). *Virtue epistemology: Essays on epistemic virtue and responsibility.* Oxford University Press on Demand.

Fanon, F. (2007). *The wretched of the earth.* Grove/Atlantic, Inc.

Fara, M. (2008). Masked abilities and compatibilism. *Mind*, *117*(468), 843–865.

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails.* University of Minnesota Press.

Flynn, F. G., Cummings, J. L., & Gornbein, J. (1991). Delusions in dementia syndromes: Investigation of behavioral and neuropsychological correlates. *The Journal of neuropsychiatry and clinical neurosciences.*

Fodor, J. A. [Jerry A]. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind* (Vol. 2). MIT press.

Foerstl, H., Almeida, O. P., Owen, A. M., Burns, A., & Howard, R. (1991). Psychiatric, neurological and medical aspects of misidentification syndromes: A review of 260 cases. *Psychological medicine*, *21 4*, 905–910.

Frances, B., & Matheson, J. (2019). Disagreement. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019). Metaphysics Research Lab, Stanford University.

Frankish, K. (2012). Delusions, levels of belief, and non-doxastic acceptances. *Neuroethics*, *5*(1), 23–27.

Fraser, R. (2020). Epistemic FOMO. *The Cambridge Humanities Review.*

Fraser, R. (2021). Narrative testimony. *Philosophical Studies*, 1–28.

Freeman, D. (2006). Delusions in the nonclinical population. *Current psychiatry reports*, *8*(3), 191–204.

Freeman, D. (2007). Suspicious minds: The psychology of persecutory delusions. *Clinical psychology review*, *27*(4), 425–457.

Freeman, D., Garety, P., & Kuipers, E. (2001). Persecutory delusions: Developing the understanding of belief maintenance and emotional distress. *Psychological medicine*, *31*(7), 1293.

Freeman, D., & Garety, P. A. [Philippa A]. (2004). *Paranoia: The psychology of persecutory delusions*. Psychology Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230–1233.

Frye, M. (1983). *The politics of reality: Essays in feminist theory*. Crossing Press.

Frymer, P., & Grumbach, J. M. (2021). Labor unions and white racial politics. *American Journal of Political Science*, *65*(1), 225–240.

Fuchs, T. (2005). Delusional mood and delusional perception–a phenomenological analysis. *Psychopathology*, *38*(3), 133–139.

Fuchs, T. (2015). The intersubjectivity of delusions. *World Psychiatry*, *14*(2), 178.

Fuchs, T. (2020). Delusion, reality, and intersubjectivity: A phenomenological and enactive analysis. *Philosophy, Psychiatry, & Psychology*, *27*(1), 61–79.

Gaag, M., Valmaggia, L., & Smit, F. (2014). The effects of individually tailored formulation-based cognitive behavioural therapy in auditory hallucinations and delusions: A meta-analysis. *Schizophrenia Research*, 30–37.

Gaertner, S. L., Dovidio, J. F., Nier, J. A., Banker, B. S., Ward, C. M., Houlette, M., & Loux, S. (2000). The common ingroup identity model for reducing intergroup bias: Progress and challenges. *Social identity processes: Trends in theory and research* (pp. 133–148). Sage Publications Ltd.

Galef, J. (2021). *The scout mindset: Why some people see things clearly and others don't.* Penguin.

Ganapini, M. B. (2020). Belief's minimal rationality. *Philosophical Studies*, 1–20.

Garcia, J. L. (1996). The heart of racism. *Journal of social philosophy*, *27*(1), 5–46.

Gardner, D. M., Baldessarini, R. J., & Waraich, P. (2005). Modern antipsychotic drugs: A critical overview. *Cmaj*, *172*(13), 1703–1711.

Garety, P. A. [Philippa A.], & Freeman, D. (1999). Cognitive approaches to delusions: A critical review of theories and evidence. *British Journal of Clinical Psychology*, *38*(2), 113–154.

Garety, P. A. [Phillip A], Kuipers, E., Fowler, D., Freeman, D., & Bebbington, P. (2001). A cognitive model of the positive symptoms of psychosis. *Psychol Med*, *31*(2), 189–195.

Gendler, T. S. (2008). Alief in action (and reaction). *Mind & Language*, *23*(5), 552–585.

Gendler, T. S. (2012). Intuition, imagination, and philosophical methodology: A summary. *Analysis*, *72*, 759–764.

Gerrans, P. (2001). Delusions as performance failures. *Cognitive Neuropsychiatry*, *6*(3).

Giessner, S. R., Ullrich, J., & van Dick, R. (2012). A social identity analysis of mergers & acquisitions. *The handbook of mergers and acquisitions*, 474–494.

Gilbert, D. (2006). *Stumbling on happiness*. Knopf.

Gilligan, C. (1993). *In a different voice: Psychological theory and women's development*. Harvard University Press.

Godfrey-Smith, P. (2005). *Folk psychology as a model*. Ann Arbor, MI: Michigan Publishing, University of Michigan Library.

Goffman, E. (1978). *The presentation of self in everyday life*. Harmondsworth London.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press.

Green, M. (2016). Expressing, showing, and representing. *The expression of emotion: Philosophical, psychological, and legal perspectives*, 25–45.

Greene, J. A. (2005). Cognitive-behavioral therapy. *Psychiatric Services*, *56*(9), 1161–1162.

Greenebaum, J., & Dexter, B. (2018). Vegan men and hybrid masculinity. *Journal of Gender Studies*, *27*(6), 637–648.

Grimm, S. R. (2016). How understanding people differs from understanding the natural world. *Philosophical Issues*, *26*(1), 209–225.

Handfield, T., & Bird, A. (2008). Dispositions, rules, and finks. *Philosophical Studies*, *140*(2), 285–298.

Hannon, M. (2018). *What's the point of knowledge?: A function-first epistemology*. Oxford University Press.

Hannon, M. (2020). Empathetic understanding and deliberative democracy. *Philosophy and Phenomenological Research*, *101*(3), 591–611.

Hannon, M. (2021). Disagreement or badmouthing? the role of expressive discourse in politics. In M. Hannon & E. Edenberg (Eds.), *Political epistemology*. Oxford University Press.

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, *99*(1999), 315–331.

Harmon-Jones, E., & Harmon-Jones, C. (2007). Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, *38*(1), 7–16.

Haslanger, S. (2011). Ideology, generics, and common ground. *Feminist metaphysics* (pp. 179–207). Springer.

Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15.

Haslanger, S. (2017). Racism, ideology, and social movements. *Res Philosophica*, *94*(1), 1–22.

Haslanger, S. (2019). Cognition as a social skill. *Australasian Philosophical Review*, *3*(1), 5–25.

Haslanger, S. (2020). Failures of methodological individualism: The materiality of social systems. *Journal of Social Philosophy*.

Haslanger, S. (2022a). How to change a social structure. In R. Chang & A. Srinivasan (Eds.), *Normative philosophy: Conversations in moral, legal, and political philosophy*. Oxford University Press.

Haslanger, S. (2022b). Ideology in practice: What does ideology do? *The Aquinas Lecture*.

Helton, G. (2020). If you can't change what you believe, you don't believe it. *Noûs*, *54*(3), 501–526.

Hofstadter, R. (2012). *The paranoid style in american politics*. Vintage.

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology*. Oxford University Press.

Hughey, M. W. (2014). White backlash in the 'post-racial'united states. *Ethnic and Racial Studies*, *37*(5), 721–730.

James, W. (1979). *The will to believe and other essays in popular philosophy* (Vol. 6). Harvard University Press.

Jaspers, K. (1963). *General psychopathology* (J. Hoenig & M. Hamilton, Trans.). University of Chicago Press.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, *68*(3), 221–263.

Jordan, C. H., Spencer, S. J., Zanna, M. P., Hoshino-Browne, E., & Correll, J. (2003). Secure and defensive high self-esteem. *Journal of personality and social psychology*, *85*(5), 969.

Jordan, H. W., & Howe, G. (1980). De Clerambault syndrome (erotomania): A review and case presentation. *Journal of the National Medical Association*, *72*(10).

Jost, J. T. (2019). A quarter century of system justification theory: Questions, answers, criticisms, and societal applications. *British Journal of Social Psychology*, *58*(2), 263–314.

Kahan, D. M. (2012). Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision making*, *8*, 407–24.

Kahan, D. M. (2015). The expressive rationality of inaccurate perceptions. *Behavioral & Brain Sciences*, *40*, 26–28.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. *SSRN Electronic Journal*.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. (2013). Presumed fair: Ironic effects of organizational diversity structures. *Journal of personality and social psychology*, *104*(3), 504.

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American journal of Psychiatry*, *160*(1), 13–23.

Kelly, T. (2008). Disagreement, dogmatism, and belief polarization. *Journal of Philosophy*, *105*(10), 611–633.

Kelly, T. (2013). Evidence can be permissive. In M. Steup & J. Turri (Eds.), *Contemporary debates in epistemology* (pp. 298–311). Blackwell.

Kendall, P., & Bemis, K. (1983). Thought and action in psychotherapy: The cognitive-behavioral approaches. *The clinical psychology handbook* (pp. 562–592). Pergamon.

Kiecolt, K. J. (2000). Self-change in social movements. *Self, identity, and social movements*, 110–131.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation* (pp. 385–418). Elsevier.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228.

Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of experimental social psychology*, *28*(2), 145–168.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, *127*(2), 242–257.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480–498.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of personality and social psychology*, *68*(4), 565.

Lacombe, M. J. (2019). The political weaponization of gun owners: The national rifle association's cultivation, dissemination, and use of a group social identity. *The Journal of Politics*, *81*(4), 1342–1356.

LaFrance, A. (2020). Q-anon is more important than you think. *The Atlantic.*

Landemore, H. (2017). *Democratic reason: Politics, collective intelligence, and the rule of the many.* Princeton University Press.

Lasonen-Aarnio, M. (2020). Perspectives and good dispositions. *Philosophy and Phenomenological Research.*

Leeuwen, N. V. (2014). Religious credence is not factual belief. *Cognition*, *133*(3), 698–715.

Lerman, C., Croyle, R., Tercyak, K., & Hamman, H. (2002). Genetic testing: Psychological aspects and implications. *Journal of Consulting and Clinical Psychology*, *70*, 784–797.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind." *Psychological review*, *94*(4), 412.

Levendusky, M. (2009). *The partisan sort.* University of Chicago Press.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*, *49*(4), 800–823.

Lewis, D. (1974). Radical interpretation. *Synthese*, *27*(July-August), 331–344.

Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly*, *47*(187), 143–158.

Lewis, D. K. (1982). Logic for equivocators. *Noûs*, *16*(3), 431–441.

Liberman, A., & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin*, *18*(6), 669–679.

Lincoln, T., & Peters, E. (2018). A systematic review and discussion of symptom specific cognitive behavioural approaches to delusions and hallucinations. *Schizophrenia Research*, *203*, 66–79.

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Lüllmann, E., & Lincoln, T. M. (2013). The effect of an educating versus normalizing approach on treatment motivation in patients presenting with delusions: An experimental investigation with analogue patients. *Schizophrenia research and treatment*, *2013*.

Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American sociological review*, 999–1017.

Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, *3*.

Madva, A. (2017). Biased against debiasing: On the role of (institutionally sponsored) self-transformation in the struggle against prejudice. *Ergo: An Open Access Journal of Philosophy*, *4*.

Madva, A. (2019). The inevitability of aiming for virtue. In B. R. Sherman & S. Goguen (Eds.), *Overcoming epistemic injustice* (pp. 85–99). Rowman; Littlefield International.

Madva, A. (2020). Individual and structural interventions. *An introduction to implicit bias* (pp. 233–270). Routledge.

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of individual psychology*, *30*(1), 98.

Maher, B. A. (1999). Anomalous experience in everyday life: Its significance for psychopathology. *The Monist*, *82*(4), 547–570.

Maibom, H. L. (2007). The presence of others. *Philosophical Studies*, *132*(2), 161–190.

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 629–658.

Mandelbaum, E. (2019). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind and Language*, *34*(2), 141–157.

Markus, G. B. (1986). Stability and change in political attitudes: Observed, recalled, and "explained". *Political behavior*, *8*(1), 21–44.

Martin, C. B. (1994). Dispositions and conditionals. *The Philosophical Quarterly (1950-)*, *44*(174), 1–8.

Martin Luther King, J. (1956). The 'New Negro' of the South: Behind the montgomery story. *Socialist Call*, *24*, 16–19.

Mason, L. (2018). *Uncivil agreement: How politics became our identity.* University of Chicago Press.

McDowell, J. (1998). Having the world in view: Sellars, Kant, and intentionality. *Journal of Philosophy*, *95*(9), 431–492.

McGeer, V. (2007a). The regulative dimension of folk psychology. *Folk psychology re-assessed* (pp. 137–156). Springer.

McGeer, V. (2007b). The regulative dimension of folk psychology. In D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Springer.

McGrath, M. (2020). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*.

McHoskey, J. W. (1995). Case closed? on the John F. Kennedy assassination: Biased assimilation of evidence and attitude polarization. *Basic & Applied Social Psychology*, *17*(3), 395–409.

McKay, R. (2012). Delusional inference. *Mind & Language*, *27*(3), 330–355.

McKay, R. P., & Cipolotti, L. (2007). Attributional style in a case of cotard delusion. *Consciousness and Cognition*, *16*, 349–359.

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination.* Oxford University Press.

Metz, C. (2021). Silicon valley's safe space. *New York Times.*

Mills, C. (2007). White ignorance. *Race and epistemologies of ignorance*, *247*, 26–31.

Miracchi, L. (2019). When evidence isn't enough: Suspension, evidentialism, and knowledge-first virtue epistemology. *Episteme*, *16*(4), 413–437.

Moreno, K. N., & Bodenhausen, G. V. (1999). Resisting stereotype change: The role of motivation and attentional capacity in defending social beliefs. *Group Processes & Intergroup Relations*, *2*(1), 5–16.

Moritz, S., Andreou, C., Schneider, B. C., Wittekind, C. E., Menon, M., Balzan, R. P., & Woodward, T. S. (2014). Sowing the seeds of doubt: A narrative review on metacognitive training in schizophrenia. *Clinical psychology review*, *34*(4), 358–366.

Moritz, S., & Woodward, T. S. (2004). Plausibility judgment in schizophrenic patients: Evidence for a liberal acceptance bias. *German Journal of Psychiatry*, *7*(4), 66–74.

Moritz, S., & Woodward, T. S. (2005). Jumping to conclusions in delusional and non-delusional schizophrenic patients. *British Journal of Clinical Psychology*, *44*(2), 193–207.

Moritz, S., & Woodward, T. S. (2006). A generalized bias against disconfirmatory evidence in schizophrenia. *Psychiatry research*, *142*(2-3), 157–165.

Moritz, S., & Woodward, T. S. (2007). Metacognitive training in schizophrenia: From basic research to knowledge translation and intervention. *Current opinion in psychiatry*, *20*(6), 619–625.

Morrell, M. E. (2010). *Empathy and democracy: Feeling, thinking, and deliberation.* Penn State Press.

Morton, J. M. (2014). Cultural code-switching: Straddling the achievement gap. *Journal of Political Philosophy*, *22*(3), 259–281.

Mullins, S., & Spence, S. A. (2003). Re-examining thought insertion: Semi-structured literature review and conceptual analysis. *The British Journal of Psychiatry*, *182*(4), 293–298.

Murdoch, I. (1956). Symposium: Vision and choice in morality. *Aristotelian Society Supplementary Volume*, *30*(1), 32–58.

Mutz, D. C. (2006). *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Myin-Germeys, I., Nicolson, N., & Delespaul, P. A. (2001). The context of delusional experiences in the daily life of patients with schizophrenia. *Psychological medicine*, *31*(3), 489.

Nguyen, C. T. [C. Thi]. (2020a). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161.

Nguyen, C. T. [C Thi]. (2020b). *Games: Agency as art*. Oxford University Press, USA.

Nguyen, C. T. [C Thi]. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplements*, *89*, 227–255.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220.

Oaksford, M., Chater, N. et al. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

O'Connor, C., & Weatherall, J. O. (2019). *The misinformation age*. Yale University Press.

Pandis, C., Agrawal, N., & Poole, N. (2019). Capgras' delusion: A systematic review of 255 published cases. *Psychopathology*, *52*, 1–13.

Pascal, B. (1852). *Pensées*. Dezobry et E. Magdeleine.

Paul, S., & Morton, J. (2018). Grit. *Ethics*, *129*(2), 175–203.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Piazza, J., Ruby, M. B., Loughnan, S., Luong, M., Kulik, J., Watkins, H. M., & Seigerman, M. (2015). Rationalizing meat consumption. the 4 Ns. *Appetite*, *91*, 114–128.

Pickard, H. (2015). Psychopathology and the ability to do otherwise. *Philosophy and Phenomenological Research*, *90*(1), 135–163.

Pickard, H. (2020). Addiction and the self. *Noûs*, *55*(4), 737–761.

Pickard, H., & Ward, L. (2013). Responsibility without blame: Philosophical reflections on clinical practice. *Oxford handbook of philosophy of psychiatry*, 1134–1154.

Porot, N., & Mandelbaum, E. (2020). The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1539.

Prior, E. W., Pargetter, R., & Jackson, F. (1982). Three theses about dispositions. *American Philosophical Quarterly*, *19*(3), 251–257.

Prior, M., & Lupia, A. (2008). Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science*, *52*(1), 169–183.

Pyszczynski, T., Greenberg, J., & Holt, K. (1985). Maintaining consistency between self-serving beliefs and available data: A bias in information evaluation. *Personality and Social Psychology Bulletin*, *11*(2), 179–190.

Pyszczynski, T., Solomon, S., & Greenberg, J. (2015). Thirty years of terror management theory: From genesis to revelation. *Advances in experimental social psychology* (pp. 1–70). Elsevier.

Quilty-Dunn, J. (2020). *Unconscious rationalization, or: How not to think about awfulness and death* [manuscript].

Quilty-Dunn, J., & Mandelbaum, E. (2018). Against dispositionalism: Belief in cognitive science. *Philosophical Studies*, *175*(9), 2353–2372.

Ramachandran, V. S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses*, *47*(5), 347–362.

Reed, M. B., & Aspinwall, L. G. (1998). Self-affirmation reduces biased processing of health-risk information. *Motivation and emotion*, *22*(2), 99–132.

Reimer, M. (2010). Only a philosopher or a madman: Impractical delusions in philosophy and psychiatry. *Philosophy, Psychiatry, and Psychology*, *17*(4), 315–328.

Rhodes, J., & Gipps, R. G. (2008). Delusions, certainty, and the background. *Philosophy, Psychiatry, & Psychology*, *15*(4), 295–310.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*(1), 52–73.

Riggle, N. (2015). Personal style and artistic style. *The Philosophical Quarterly*, *65*(261), 711–731.

Robinson, J. M. (1985). Style and personality in the literary work. *The Philosophical Review*, *94*(2), 227–247.

Rooney, P. (1991). Gendered reason: Sex metaphor and conceptions of reason. *Hypatia*, *6*(2), 77–103.

Rothgerber, H. (2013). Real men don't eat (vegetable) quiche: Masculinity and the justification of meat consumption. *Psychology of Men & Masculinity*, *14*(4), 363.

Rovane, C. (2019). *The bounds of agency: An essay in revisionary metaphysics*. Princeton University Press.

Rutjens, B. T., Sutton, R. M., & van der Lee, R. (2018). Not all skepticism is equal: Exploring the ideological antecedents of science acceptance and rejection. *Personality and Social Psychology Bulletin*, *44*(3), 384–405.

Ryle, G. (1949). *The concept of mind*. Hutchinson & Co.

Sankaran, K. (2020). What's new in the new ideology critique? *Philosophical Studies*, *177*(5), 1441–1462.

Sarin, F., Wallin, L., & Widerlöv, B. (2011). Cognitive behavior therapy for schizophrenia: A meta-analytical review of randomized controlled trials. *Nordic journal of psychiatry*, *65*(3), 162–174.

Sass, L., Parnas, J., & Zahavi, D. (2011). Phenomenological psychopathology and schizophrenia: Contemporary approaches and misunderstandings. *Philosophy, Psychiatry, & Psychology*, *18*(1), 1–23.

Sass, L. A. [Louis A.]. (1994). *The paradoxes of delusion: Wittgenstein, schreber, and the schizophrenic mind*. Cornell University Press.

Sass, L. A. [Louis A], & Pienkos, E. (2013). Delusion: The phenomenological approach. *KWM Fulford*.

Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? what an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly*, *82*(1), 135–147.

Schaller, M., & Cialdini, R. B. (1990). Happiness, sadness, and helping: A motivational integration. In E. T. Higgins & R. M. Sorrentino (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (pp. 265–296). The Guilford Press.

Schellenberg, S. (2018). *The unity of perception: Content, consciousness, evidence*. Oxford University Press.

Schroeter, L., & Schroeter, F. (2015). Rationalizing self-interpretation. *The palgrave handbook of philosophical methods* (pp. 419–447). Springer.

Schultheis, G. (2018). Living on the edge: Against epistemic permissivism. *Mind*, *127*(507), 863–879.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, *36*(2), 249–275.

Schwitzgebel, E. (2021). The pragmatic metaphysics of belief. In C. Borgoni, D. Kindermann, & A. Onofri (Eds.), *The fragmented mind*. Oxford University Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, *1*(19), 253–329.

Shah, N. (2003). How truth governs belief. *The Philosophical Review*, *112*(4), 447–482.

Shah, N., & Velleman, J. D. (2005). Doxastic deliberation. *The Philosophical Review*, *114*(4), 497–534.

Shelby, T. (2003). Ideology, racism, and critical social theory. *The Philosophical Forum*, *34*(2), 153–188.

Shelby, T. (2017). *Dark ghettos*. Harvard University Press.

Sherman, D. A., Nelson, L. D., & Steele, C. M. (2000). Do messages about health risks threaten the self? *Personality and Social Psychology Bulletin*, *26*(9), 1046–1058.

Sherman, D. K., & Cohen, G. L. (2002). Accepting threatening information: Self–affirmation and the reduction of defensive biases. *Current directions in psychological science*, *11*(4), 119–123.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in experimental social psychology*, *38*, 183–242.

Silva, J., Leong, G., Weinstock, R., & Boyer, C. (1989). Capgras syndrome and dangerousness. *Journal of the Americal Academy of Psychiatry and the Law Online*, *17*(1), 5–14.

Slusher, M. P., & Anderson, C. A. (1989). Belief perseverance and self-defeating behavior. *Self-defeating behaviors* (pp. 11–40). Springer.

Smith, A. D. (1977). Dispositional properties. *Mind*, *86*(343), 439–445.

Smith, M. (2003). Rational capacities, or: How to distinguish recklessness, weakness, and compulsion. In S. Stroud & C. Tappolet (Eds.), *Weakness of will and practical irrationality* (pp. 17–38). Oxford: Clarendon.

Sontag, S. (1966). On style. *Against interpretation and other essays*, *29*.

Sosa, E. (2015). *Judgment & agency*. Oxford University Press UK.

Srinivasan, A. (2020). Radical externalism. *Philosophical Review*, *129*(3), 395–431.

Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5/6), 701–721.

Stanley, J. (2015). *How propaganda works*. Princeton University Press.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. *Advances in experimental social psychology*, *21*(2), 261–302.

Steele, C. M., Spencer, S. J., & Lynch, M. (1993). Self-image resilience and dissonance: The role of affirmational resources. *Journal of personality and social psychology*, *64*(6), 885.

Stephens, G. L., & Graham, G. (2004). Reconceiving delusions. *International Review of Psychiatry*, *16*, 236–241.

Stone, T., & Young, A. W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, *12*(3–4), 327–364.

Strawson, P. (1962). Freedom and resentment. *Proceedings of the british academy, volume 48: 1962* (pp. 1–25).

Sullivan, S. (2006). *Revealing whiteness: The unconscious habits of racial privilege*. Indiana University Press.

Swann Jr, W. B. (1992). Seeking "truth," finding despair: Some unhappy consequences of a negative self-concept. *Current Directions in Psychological Science*, *1*(1), 15–18.

Taber, C., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769.

Táíwò, O. (2017). Beware of schools bearing gifts. *Public Affairs Quarterly*, *31*(1), 1–18.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific american*, *223*(5), 96–103.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, *33*(1), 1–39.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984–16.

Tannen, D. et al. (2005). *Conversational style: Analyzing talk among friends*. Oxford University Press.

Tarlow, E. M., & Haaga, D. A. (1996). Negative self-concept: Specificity to depressive symptoms and relation to positive and negative affectivity. *Journal of Research in Personality*, *30*(1), 120–127.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Tilly, C., & Tarrow, S. G. (2015). *Contentious politics*. Oxford University Press.

Todd, J., & Dewhurst, K. (1955). The Othello syndrome: A study in the psychopathology of sexual jealousy. *The Journal of nervous and mental disease*, *122*(4), 367–374.

Turkington, D., Kingdon, D., & Weiden, P. J. (2006). Cognitive behavior therapy for schizophrenia. *American Journal of Psychiatry*, *163*(3), 365–373.

Turnbull, O. H., Jenkins, S., & Rowley, M. L. (2004). The pleasantness of false beliefs: An emotion-based account of confabulation. *Neuropsychoanalysis*, *6*(1), 5–16.

Turner, J. C. (2010). Social categorization and the self-concept: A social cognitive theory of group behavior. In T. Postmes & N. R. Branscombe (Eds.), *Rediscovering social identity* (pp. 243–272). Psychology Press.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

Turner, J. C., & Oakes, P. J. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, *25*(3), 237–252.

Turner, J. C., Oakes, P. J., Haslam, S. A., & McGarty, C. (1994). Self and collective: Cognition and social context. *Personality and social psychology bulletin*, *20*(5), 454–463.

Turner, J. C., & Reynolds, K. J. (2011). Self-categorization theory. *Handbook of theories in social psychology*, *2*(1), 399–417.

Turner, M., & Coltheart, M. [M.]. (2010). Confabulation and delusion: A common monitoring framework. *Cognitive Neuropsychiatry*, *15*, 346–376.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Valdesolo, P., & DeSteno, D. (2008). The duality of virtue: Deconstructing the moral hypocrite. *Journal of Experimental Social Psychology*, *44*(5), 1334–1338.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences*, *22*(3), 213–224.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Velleman, D. (2000). On the aim of belief. In D. Velleman (Ed.), *The possibility of practical reason* (pp. 244–81). Oxford University Press.

Wang, E. W. (2019). *The collected schizophrenias*. Graywolf Press.

Wessely, S., Buchanan, A., Cutting, J., Everitt, B., Garety, P., & Taylor, P. (1993). Acting on delusions. i: Prevalence. *The British journal of psychiatry : the journal of mental science*, *163*, 69–76.

White, R. (2013). Evidence cannot be permissive. In M. Steup & J. Turri (Eds.), *Contemporary debates in epistemology* (pp. 312–323). Blackwell.

Willard-Kyle, C. (2017). Do great minds really think alike? *Synthese*, *194*(3).

Williams, B. (1970). Deciding to believe. In B. Williams (Ed.), *Problems of the self* (pp. 136–51). Cambridge University Press.

Williams, D. (2021). Socially adaptive belief. *Mind & Language*, *36*(3), 333–354.

Williamson, T. (2002). *Knowledge and its limits*. Oxford University Press on Demand.

Wollheim, R. (1987). Pictorial style: Two views. In B. Lang (Ed.), *The concept of style*. Cornell University Press.

Woodward, T. S., Buchy, L., Moritz, S., & Liotti, M. (2007). A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophrenia bulletin*, *33*(4), 1023–1028.

Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (bade) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, *28*(4), 605–617.

Woodward, T. S., Moritz, S., Menon, M., & Klinge, R. (2008). Belief inflexibility in schizophrenia. *Cognitive neuropsychiatry*, *13*(3), 267–277.

Wykes, T., Steel, C., Everitt, B., & Tarrier, N. (2008). Cognitive behavior therapy for schizophrenia: Effect sizes, clinical models, and methodological rigor. *Schizophrenia bulletin*, *34*(3), 523–537.

Young, A. [A.], & Leafhead, K. (1996). Betwixt life and death: Case studies of the cotard delusion. In P. Halligan & J. Marshall (Eds.), *Method in madness: Case studies in cognitive neuropsychiatry*. Psychology Press.

Young, A. W., Robertson, I. H., Hellawell, D., De Pauw, K., & Pentland, B. (1992). Cotard delusion after brain injury. *Psychological Medicine*, *22*(3), 799–804.

Young, I. M. (2014). Five faces of oppression. *Rethinking power*, 174–195.

Zagzebski, L. T. (1996). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Zheng, R. (2018). Bias, structure, and injustice: A reply to Haslanger. *Feminist Philosophy Quarterly*, *4*(1).

Zimmerman, A. Z. (2018). *Belief: A pragmatic picture*. Oxford University Press.