ENDORSEMENT AND INQUIRY

BY WILLIAM FLEISHER

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
Ernest Sosa and Branden Fitelson
and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2018

ABSTRACT OF THE DISSERTATION

Endorsement and Inquiry

by William Fleisher

Dissertation Directors: Ernest Sosa and Branden Fitelson

This dissertation is about the epistemology of inquiry. It concerns the appropriate attitudes of inquiry, what epistemic reasons and values such attitudes should be sensitive to, and what kinds of virtues an inquirer should have. I argue that our theories about these attitudes, reasons, and virtues must be revised. This is necessary in light of a number of problems that the practice of contemporary, organized research raises for traditional epistemological theories. Standard accounts of epistemic rationality and justification fail to properly evaluate contemporary researchers, instead counting them as irrational in their commitments, theory choice, pursuit strategies, evidential evaluations and assertions. This is an unacceptable result: our best epistemic theories should not call our best researchers epistemically irrational for promoting healthy inquiry. We need a new account.

My project seeks to resolve some of these problems in the epistemology of inquiry. First, I propose that we recognize the attitude of *endorsement*: a distinct doxastic attitude, governed by an inclusive kind of epistemic rationality. This is the attitude of resilient commitment and advocacy that researchers have toward their favored theories during inquiry. Recognizing endorsement, along with the inclusive epistemic rationality that governs it, allows us to vindicate the committed advocacy of researchers. This project is the focus of my first two chapters, in which I argue that endorsement

eases the tension between individual and collective rationality, promotes beneficial disagreement and distribution of cognitive labor, and solves the biggest problems for conciliationism about disagreement.

The third chapter takes up a related project concerning our understanding of the credences (degrees of belief) of researchers engaged in inquiry. There I argue that (like endorsement), the credences of researchers should be understood as *fragmented*. That is, these attitudes are compartmentalized to individual projects in a subject's mental life. The final chapter concerns the relationship between virtues of inquiry and virtues that constitute knowledge. This project is another attempt to alleviate tensions between our traditional epistemic theories and the practice of inquiry. I argue that the intellectual character virtues investigated by virtue responsibilists are virtues of inquiry. They gain their epistemic import by promoting healthy inquiry, in a way that supports the reliabilist virtues (or competences) that constitute knowledge. This removes the apparent conflict between the two factions of virtue epistemology, while at the same time explaining the epistemic value of intellectual character virtues by appeal to their role in inquiry, and their indirect connection to knowledge.

This dissertation is written on the new, multiple-paper model. That is, it consists in a collection of independent papers in the epistemology of inquiry. They are thematically linked, but I have not altered them significantly from their individual forms. They can thus each be read independently.

Preface

This dissertation is written on the new, multiple-paper model. That is, it consists in a collection of independent papers in the epistemology of inquiry (along with an introduction). These papers are thematically linked, but I have not altered them significantly from their individual forms. They can thus each be read independently. Two of the papers are now published:

1. Chapter Two, "Rational Endorsement," is forthcoming in *Philosophical Studies* (https://doi-org.proxy.libraries.rutgers.edu/10.1007/s11098-017-0976-4)

2. Chapter Five, "Virtuous Distinctions," is published in *Synthese* (194 (8):2973–3003 (2017)).

Acknowledgements

I first want to express my deep gratitude to my committee co-chairs, Ernest Sosa and Branden Fitelson, who have been exceptionally helpful and supportive. Second, I want to thank my committee members Susanna Schellenberg and Andy Egan, who have also been extremely supportive and generous with time and feedback. Third, I want to thank Adam Elga for many years of discussion on the topics of this dissertation.

I also want to thank Howard McGary for years of guidance and support.

For written comments, and untold hours of helpful discussion, thank you to my friends, colleagues, and teachers, including: Sara Aronowitz, Ben Bronner, Liz Camp, Sam Carter, Eddy Chen, Matt Duncan, Frankie Egan, Peter van Elswyk, Carolina Flores, Elizabeth Fricker, Alvin Goldman, Veronica Gomez, Jimmy Goodrich, E. J. Green, Chris Hauser, Alan Hájek, Caley Howland, Tyler John, Anton Johnson, Nico Kirk-Giannini, Zoe Johnson King, Steph Leary, Ting-An Lin, Ricardo Mena, Brian McLaughlin, Olivia Odoffin, Dee Payton, Daniel Rubio, Jonathan Schaffer, Joshua Schecter, Eli Shupe, Nick Tourville, Isaac Wilhelm, and anyone I'm forgetting.

Special thanks to my friends who took part in the Sosa dissertation workshop, for reading my papers over and over (and over) again: Austin Baker, Bob Beddor, David Black, Danny Forman, Georgi Gardiner, Lisa Miracchi, Laura Callahan, Marilie Coetsee, Pamela Robinson, Kurt Sylvan, Chris Willard-Kyle.

Thanks, also, to my parents, Susan Fleisher and Glenn Fleisher, and my brother James Fleisher, for all the support I needed to make it this far.

Finally, my deepest gratitude to Megan Feeney, without whom there is no way this dissertation would exist.

Dedication

To my family, and to Megan Feeney.

Table of Contents

Chapter 1

Introduction

Researchers are willing to assert their own theories' superiority, defend their views doggedly, repeatedly advocate them in print and in conversation, and remain committed to them in the face of difficult objections and seeming counterexamples. This leads to a puzzle: In everyday contexts, such behavior is only justified when a subject believes, or knows, the claims she advocates. But given the pervasive disagreement in cutting-edge research fields, researchers would be irrational to believe their theories. Yet, surely our best researchers are not epistemically irrational or irresponsible for being committed advocates of their views. On the contrary, such behavior is constitutive of healthy inquiry.

This puzzle places our overall account of epistemic rationality in danger of serious incoherence, which would leave us unable to effectively evaluate the performance of researchers. Yet we have been dragged to this position by two seemingly persuasive lines of argument. On the one hand, epistemologists and philosophers of language have argued that warranted assertion requires a high epistemic standard (DeRose 2016; Goldberg 2015; Schaffer 2008; Williamson 2000). This claim also has significant empirical evidence in its favor (Turri 2017). On the other hand, social epistemologists, philosophers of science, and psychologists have described the difficulty of scientific inquiry, and argued for the importance of having researchers pursuing a variety of options (Kitcher 1990; L. Laudan 1978; Strevens 2003; Zollman 2010). This last claim about the division of labor also has significant empirical support (De Cruz & De Smedt 2013; Mercier 2016; Mercier & Sperber 2011). We are thus led to the result that researchers who serve as committed advocates of their own theories—a practice we know to be beneficial to inquiry—are thereby behaving (epistemically) irrationally.

Inquiry is a paradigmatic epistemic endeavor. That we gain knowledge through inquiry is a platitude. This makes it worrisome that our traditional epistemological theories seem to evaluate excellent researchers as irrational (or unjustified). Researchers' committed advocacy of their theories, however, is only one example of this kind of problem for inquiry in light of our traditional epistemological views.

My project seeks to resolve some of these problems in the epistemology of inquiry. First, I propose that we recognize the attitude of *endorsement*: a distinct doxastic attitude, governed by an inclusive kind of epistemic rationality. Recognizing endorsement allows us to vindicate the committed advocacy of researchers. This project is the focus of my first two chapters. The third chapter takes up a related project concerning our understanding of the credences (degrees of belief) of researchers engaged in inquiry. There I argue that (like endorsement), the credences of researchers should be understood as *fragmented*. The final chapter concerns the relationship between virtues of inquiry and virtues that constitute knowledge. This project is another attempt to alleviate tensions between our traditional epistemic theories and the practice of inquiry. I argue that the intellectual character virtues investigated by virtue responsibilists gain their epistemic import by having the right kind of connection to the virtues (or competences) that constitute knowledge. This removes the apparent conflict between the two factions of virtue epistemology, while at the same time explaining the epistemic value of intellectual character virtues by their connection to knowledge.

This dissertation is written on the new, multiple-paper model. That is, it consists in a collection of independent papers in the epistemology of inquiry. They are thematically linked, but I have not altered them significantly from their individual forms. They can thus each be read independently. Two of the papers are now published: Chapter One, "Rational Endorsement," is forthcoming in *Philosophical Studies*. Chapter Two, "Virtuous Distinctions," is published in *Synthese* (194 (8):2973–3003 (2017)). The published versions do not differ significantly from what is written here.

In the rest of this introduction, I will give a brief overview of each of these projects. In my first two chapters, I present the theory of endorsement. Endorsement is the attitude of resilient, committed advocacy that an inquirer takes toward her favored

theory. Rational endorsement has different epistemic standards than belief because it is rationally sensitive to considerations beyond evidence for the theory in question. Endorsement is appropriately sensitive to reasons concerning the collective goals of inquiry. Thus, it is rational to endorse a theory when doing so promotes the health of inquiry in general. As a result, one can endorse a theory when one is less confident in the theory than in its negation, or even when one is more confident in some other theory. This is because a researcher's considerations about contributing to the goals of collective inquiry can outweigh her evidence against a theory.

Recognizing endorsement allows us to vindicate the actions of at least some inquirers in cutting-edge fields. It provides a proper account of many researchers who are genuinely committed to the goals of inquiry, and who therefore pursue theories that are promising but unlikely. The theory of rational endorsement also prescribes several ways in which our practices of inquiry should be reformed. In addition, endorsement helps solve a variety of more specific problems, including in social epistemology and the general philosophy of science.

In the first chapter, I apply the theory of endorsement to social epistemology, specifically, to easing tensions between collective and individual epistemic rationality. In pursuit of this, I develop a notion of *inclusive* epistemic rationality, which governs rational endorsement. This kind of rationality involves both familiar *intrinsic* epistemic reasons, and those I call *extrinsic*. Intrinsic epistemic reasons are reasons to think a proposition is true, that indicate its truth. Extrinsic reasons, in contrast, are epistemic considerations in favor of endorsing a proposition that have to do with the overall health and longterm success of inquiry. Rational beliefs (and degrees of belief) are insensitive to such extrinsic concerns. But endorsement is sensitive to them, as what it is rational to endorse depends on what will contribute to successful inquiry.

One such extrinsic reason has to do with distribution of cognitive labor. That a theory has too few researchers working on it can serve as a genuinely good reason to pursue and advocate for a it: we don't want to put all our epistemic eggs in a single theoretical basket. This healthy commitment to, pursuit of, and advocacy for, less-likely theories is explained and justified by endorsement. Thus, this treatment helps

ease a particular tension between individual and collective rationality that arises when dealing with the distribution of cognitive labor: it justifies pursuit of theories which are not the most likely to be true, but which are important possibilities to be explored. To explain this solution, I develop a formal decision theory for rational endorsement, which characterizes inclusive epistemic rationality. This application of endorsement comprises my first chapter.

A second application of the endorsement project, which I pursue in the second chapter, is to the epistemology of disagreement. Specifically, I argue that endorsement can solve two significant problems for *conciliationism*.

The following principle seems highly intuitive to many philosophers: when we encounter genuine disagreement with a peer who we take to be equally reliable about a particular question, we should (at least often) lower our confidence in our own answer to that question. Theories that embrace this principle are called conciliationist views. Conciliationism is a popular and plausible view about how to respond to disagreement with one's peers. However, it is beset by two difficult problems: the skeptical results problem and the self-undermining problem. These problems are purported to be strong enough to force proponents to give up conciliationism. The first problem is that, given the systematic disagreement in cutting-edge research domains, conciliationism calls for much lower confidence in theories than researchers seem to display in their behavior. The second problem is that conciliationism, when applied to the epistemology of disagreement, is self-undermining: adopting the view is incompatible with being justified in adopting the view.

Endorsement provides a simple response to both problems. The skeptical results problem is resolved by appealing to endorsement as the right kind of committed advocacy for researchers to have. This means that researchers' lack of justified beliefs in their theories does not lead to impractically agnostic results. The committed advocacy we wish to account for is explained by endorsement, not belief.

Endorsement solves the self-undermining problem in two steps. First, conciliatory views have (appropriately) been characterized in terms of how a subject's beliefs (or degrees of belief) should be updated or modulated in response to disagreement.

Such theories do not entail any requirement with respect to endorsement, as well they should not. And since disagreement about disagreement cases occur within a cutting-edge research domain, endorsement is the rational attitude to take towards one's theory in such cases. Second, the theory of endorsement provide a strong, principled argument against the specific kind of enkratic principles that the self-undermining argument presupposes. These are principles that connect a subject's beliefs (or credences) about theories, with whether they should follow the rules contained in those theories. So, no undermining or inconsistency arises when philosophers merely endorse their conciliatory theory.

In my third chapter, I take up a different problem for the epistemology of inquiry, concerning a different kind of inquiring attitude, specifically credences. Here I apply the notion of fragmentation, one of the same tools from the theory of endorsement. Treating credences (or degrees of belief) as fragmented helps to solve *The Problem of Old Evidence* for Bayesian Confirmation Theory. This is another problem caused by a failure of our traditional theories (here a traditional theory in formal epistemology) to accommodate aspects of inquiry. In this case, it is BCT's linear view of evidence accumulation modeled on simple perceptual learning in individual cases that is causing the problem.

Bayesian Confirmation Theory is our best formal framework for describing inductive reasoning. The Problem of Old Evidence is a particularly difficult one for confirmation theory, because it suggests that this framework fails to account for central and important cases of inductive reasoning and scientific inference. I show that we can appeal to the fragmentation of doxastic states to solve this problem for confirmation theory. This fragmented solution is independently well-motivated because of the success of fragmentation in solving other problems. Recognizing that belief states can be fragmented allows us to successfully apply the theory to the problem cases without significant changes to the simple and intuitive formalism of BCT. I also argue that the fragmentation solution is preferable to other solutions to the Problem of Old Evidence. These other solutions are already committed to something like fragmentation, but suffer from difficulties due to their additional commitments. If these

arguments are successful, Bayesian Confirmation Theory is saved from the Problem of Old Evidence, and the argument for fragmentation is bolstered by its ability to solve yet another problem.

My final chapter concerns virtue epistemology, and distinct kinds of virtues. I argue that recognizing better distinctions between kinds of virtues, including discovery-relevant virtues of inquiry, will ease the tension between different kinds of virtue epistemologists.

Virtue epistemology has been divided into two camps: *reliabilists* and *responsibilists*. This division has been attributed in part to a focus on different types of virtues, namely, faculty virtues and character virtues. I argue that this distinction is unhelpful, and that we should carve up the theoretical terrain differently. Instead, we should recognize different distinctions among virtues. These distinctions do more than help with the debate between the factions. They also help to shed light on the differences between virtuous collective inquiry and more traditional cases of learning, and help to explain the connections between the two sorts of circumstances.

I propose three distinctions among virtues (or competences; I will use the terms interchangeably). The first is a re-interpretation of Sosa's (2011) distinction between constitutive and auxiliary competences. Constitutive competences are those whose exercise can directly constitute knowledge, while auxiliary competences are those that develop, support, or enable constitutive competences. Second, I distinguish between discovery-relevant and justificatory virtues. Discovery virtues are competences to engage successfully in inquiry. For instance, they might be competences to design experiments, create new theories and hypotheses, or competences to pursue and advocate for theories. Justificatory competences are those which directly enable or support knowledge (including constitutive virtues). These would include competences to put one in a position to know, sometimes by successfully deploying constitutive competences. Finally, I distinguish between singular competences, which we can narrowly define within specific environmental conditions, and sets of competences which are composed of the singular competences.

Making these distinctions among virtues sheds light on several issues. First, it

highlights that responsibilists and reliabilists are actually engaged in different, complementary projects. Second, it shows that certain responsibilist critiques of reliabilism miss the mark. Most importantly, it explains the relationship between virtues of inquiry and the more traditional focus of epistemology: knowledge.

The main upshot is that intellectual character virtues, such as open-mindedness and intellectual courage, must involve collective auxiliary competences, including discovery-relevant ones. That is, possession of a virtue like intellectual courage requires possession of large set of individual auxiliary competences. What connects the set is a family-resemblance involving appropriate willingness to continue in the face of significant danger. Virtues in the set will include discovery virtues: for example, one to create and run important experiments (rather than safely publishable ones) in the face of risk of losing tenure. These are virtues that promote healthy inquiry. They support and enable constitutive virtues (that deliver knowledge), but largely indirectly.

The picture that emerges is that many intellectual character virtues are virtues of inquiry. They involve sensitivity to promoting healthy inquiry, and thus a sensitivity to extrinsic epistemic value. (In future work, I hope to draw additional connections between extrinsic epistemic reasons and discovery virtues.) The responsibilist project, focusing as it does on intellectual character virtues (i.e., collective auxiliary competences), is a project in the epistemology of inquiry. Virtue reliabilists, on the other hand, have been concerned with knowledge.

With this picture on the table, we can see that the virtue responsibilist project importantly depends on the reliabilist project. I argue that the distinctively epistemic value of the responsibilist's character virtues is derived from their connections to the reliabilist's constitutive virtues. This provides a unified account of the epistemic value of intellectual character virtues, one that treats them as virtues of inquiry that are indirectly related to knowledge, in a manner familiar from the previous discussion of extrinsic epistemic value.

Chapter 2

Rational Endorsement

Forthcoming in *Philosophical Studies*

2.1 Introduction

Consider the following case:

> Ellie is an entomologist, studying the brilliantly colored Madagascan Sunset Moth. Somewhat unusually, the coloration on this moth's wings are not the result of pigmentation, but of complicated light interference and polarization patterns caused by the micro-structure of its wings. The common view among Ellie's colleagues is that the function of this coloration is aposematic, i.e., to warn potential predators of the moth's toxicity.
>
> However, Ellie is aware of a new hypothesis, that the color and polarized light patterns are actually a signaling method between moths. She is not confident in the truth of this hypothesis, as no behavioral studies have yet been done. In fact, she is more confident in the aposematic theory. However, since few entomologists are working on the signaling theory, she decides to pursue it.
>
> As she garners evidence for the signaling theory, she advocates for it in published work and in discussion, defending it from various objections. She sometimes asserts its truth, or defends it as the best view. It becomes her favored theory to defend, and she remains committed to it in her work over time, even though most of her colleagues disagree with her. However, outside of professional activities, Ellie will admit that she still does not believe it. She is not that confident it is right.[1]

Ellie's situation is a familiar one: researchers often find themselves committed to theories in the face of disagreement with their peers. Moreover, it seems clear that Ellie is not behaving irrationally. Indeed, her actions are consistent with being an excellent researcher. A healthy community of inquiry, in any field, needs researchers

[1] This example is inspired by the actual case of the Madagascan Sunset Moth, *Chrysiridia rhipheus*. For details on the signaling hypothesis, see Yoshioka and Kinoshita (2007).

who will strike out to pursue theories which are not currently the most probable, or best-confirmed. Inquiry needs researchers like Ellie.

Ellie does not believe that the signaling theory is true.[2] Nor, I would suggest, should she believe it. Nonetheless, she is committed to the view, defends it, advocates for it, directs her future research on the basis of it, and even asserts that it is the right theory. These actions are generally associated with beliefs, and with epistemic standards which Ellie clearly does not meet. She does not know the signaling theory is true, so she neither obeys the knowledge norm of assertion, nor does she base her actions only on premises she knows.[3] So how can we account for the apparent rationality of her behavior?

In order to explain and justify Ellie's actions, we must recognize a distinct attitude that researchers have toward their theories. I call this attitude *endorsement*. As I will argue, endorsement is the rational attitude to take toward one's favored theory during the course of inquiry.

Endorsement is governed by what I call *inclusive* epistemic rationality. This is distinct from the veritistic (or accuracy-first) epistemic rationality that plausibly governs full belief and credence (Goldman 1986; Joyce 1998; Pettigrew 2016). It is also distinct from a true pragmatist view, which denies any distinction between pragmatic and epistemic rationality (Rinard 2015). This notion of inclusive rationality is necessary to account for the actions of excellent researchers like Ellie. Sometimes researchers are sensitive to reasons which speak in favor of pursuing a theory, but which are not reasons to think the theory is true. Yet these are not simply pragmatic reasons: they are considerations about what is good for inquiry. I will call such considerations *extrinsic* epistemic reasons (following Steel 2010).

In this paper, I will argue that recognizing endorsement is necessary in order to provide a proper account of inquiry, and the norms that govern it. The paper has two

---

[2]Nor does she believe that it is approximately enough true, or that it is the best theory, or even that it is empirically adequate in van Fraassen's sense (1980).

[3]One can even swap out "knows" here for "is justified in believing," and the knowledge norm for some other norm. Plausibly, Ellie will also fail to meet such weakened requirements. For the knowledge norm of assertion, see Williamson (1996, 2000). For an overview of the norms of assertion literature, see Pagin (2016); Weiner (2007).

goals. The first goal is to characterize when it is epistemically rational to endorse a theory. The second goal is to show that the right account of rational endorsement will allow us to smooth some apparent tensions between individual and collective rationality. The successful completion of these two projects shows that a theory of inquiry which includes endorsement provides a better account of inquiry, and prescribes better norms for inquiry, than one which does not recognize the attitude.

In what follows, I will briefly give an account of the nature of endorsement. Then, I will describe the valuable features of inquiry to which endorsement contributes (§2.3). In §2.4, I will give an account of inclusive epistemic rationality. Then, in §2.5, I will discuss some extrinsic epistemic reasons derived from the social epistemology literature regarding the distribution of cognitive labor. In §2.6.1, I argue that endorsement allows us to account for the way individual researchers are sensitive to these extrinsic reasons, and thereby helps smooth a tension between individual and collective rationality. Finally, I will provide a formal decision-theoretic framework for describing rational endorsement which reflects the foregoing considerations (§2.6.2).

2.2 Endorsement

Endorsement is a propositional, doxastic attitude; i.e., it is an attitude one takes toward a proposition. It is the appropriate attitude to take toward one's favored theory during the course of inquiry, within the domain of a cutting-edge research field.

This notion of endorsement is inspired by work on the distinction between acceptance and belief, especially by the work of L. Jonathan Cohen (1989a, 1989b, 1995), Isaac Levi (1974, 1980, 2004a), Patrick Maher (1993), and Bas van Fraassen (1980). However, there are a dizzying array of different notions of acceptance[4] and my concept of endorsement is distinct from these previous ideas of acceptance in a variety of

[4]There are at least five kinds of "acceptance" notions that appear in philosophy: Cohen's notion from epistemology (Alston 1996; Cohen 1989a), the notion from the philosophy of language (R. C. Stalnaker 1987), the notion from the philosophy of science (Kaplan 1981a, 1981b; Levi 1974; Maher 1993; Van Fraassen 1980), the concept of acceptance from the metacognition literature (Frankish 2004; Proust 2013), and the genus conception of acceptance (Shah & Velleman 2005). For more on the various notions of acceptance, see McKaughan (2007).

ways. Most importantly, endorsement is distinctively epistemic and provisional.[5] It is sensitive to both intrinsic and extrinsic epistemic considerations (see section 2.3 below for more on this distinction). It is an attitude one takes during the "context of pursuit" (L. Laudan 1978), and is importantly involved with planning future research.[6]

The following is a characterization of endorsement. It describes a set of features of the attitude which serve to distinguish endorsement from other kinds of mental states, including belief. I take it that this characterization is compatible with any particular view of the nature of mental states, with the possible exception of eliminativist views.

> **Endorsement:** Endorsement is a doxastic propositional attitude. S endorses p in a research domain d only if:
>
> 1. S is disposed to assert that p, or otherwise express commitment to p (in d).
> 2. S takes herself to be obligated to defend p (in d).
> 3. S treats p as a premise in her further reasoning (in d).
> 4. S shapes her research program in d (in part) based on p.
> 5. S is resiliently committed to p (in d).
> 6. S takes p to be a live option (i.e., she does not know p is false).
> 7. In endorsing p, S aims to promote healthy inquiry.

Endorsement is specific to a research domain: a subject endorses something for the purposes of a particular domain of inquiry. Characterizing these domains might be tricky in some cases, but I am conceiving of them as standard, familiar divisions

---

[5] I am not the first to notice the need for a provisional, acceptance-like attitude. Goldberg notices this need in the context of pervasive disagreement in philosophy (2013a; 2013b). Other examples of varying degrees of similarity can be found in Firth (1981); Lacey (2015); R. Laudan (1987); McKaughan (2007); Whitt (1985, 1990). I take this convergence to be good evidence in favor of the existence of the attitude I call endorsement. However, my account is significantly divergent from prior accounts, and I apply the theory to both philosophy and science.

[6] Following Laudan, a number of philosophers of science have appealed to this notion of the context of pursuit. This context is the stage of inquiry when researchers pursue promising but as-yet unconfirmed theories. For an overview of this literature, see McKaughan (2007) and Whitt (1990). Other examples include R. Laudan (1987); McKaughan (2007, 2008); McMullin (1976); Nickles (1981); Šešelja, Kosolosky, and Straßer (2012); Šešelja and Straßer (2013, 2014); Whitt (1985, 1992). There is insufficient space here to show all of the applications of endorsement, and inclusive epistemic rationality, to the pursuit literature.

of inquiry: fields and sub-fields of research (e.g., physics, or epistemology). Which theories are candidates for endorsement depends on the particular domain of inquiry.[7]

Endorsement is a "fragmented" attitude, meaning that it is compartmentalized rather than being a global feature of the subject's mental state. It is fragmented in two respects: first, it explains the seemingly inconsistent actions of subjects who are committed to a position during research, but are not confident enough to believe it or act on it outside of the research domain. And second, what one endorses in one domain can be inconsistent with what one endorses in a different research domain.[8]

Endorsement is similar to belief in a number of ways. However, the above characterization provides for distinguishing the two attitudes, given appropriate interpretation of its conditions.

First, condition 5 requires resiliency, as endorsement is more resilient than belief. Endorsement can be permissibly maintained in the face of significant contrary evidence, objections, or purported counter-examples. If one discovers such evidence against a proposition one believes, one should give up that belief. But endorsement is an attitude suitable even in the face of pervasive disagreement and significant contrary evidence. It is the attitude a committed advocate has toward a theory. In this way, endorsement licenses maintaining healthy disagreement. The importance of this will become clear in §2.3.[9]

A second way endorsement is distinguished from belief is the obligation to defend (condition 2). Quite often, subjects will not be required to defend their beliefs (though of course in other cases they will be). But taking up an endorsement attitude will almost invariably involve some obligation to defend the view. After all, taking a position

---

[7]Endorsement is a propositional attitude. However, this does not mean that the theories being endorsed need to be understood as propositions. I want to remain neutral on the nature and structure of scientific theories (Frigg & Nguyen 2016; Winther 2016). Technically speaking, the propositions that the attitude is taken toward can be propositions about the theory. So, if one has the view that, for instance, theories are actually models, then the proposition one endorses could be just "Theory A is an accurate enough model" or "Theory A is the best model," or some other variation on these lines.

[8]For more on the notion of fragmentation, see Egan (2008a); Elga and Rayo (2015); Lewis (1982); Rayo (2013); R. C. Stalnaker (1987).

[9]Note that this resiliency is quite distinct from Leitgeb's notion of stability. Stability involves having good reason to expect no evidence that would warrant giving up the belief (2014). Belief is stable but not resilient, and endorsement is resilient but not stable.

and advocating for it in a field of research is to be involved in defending the view.[10] One way endorsement is valuable for inquiry is because it motivates and licenses such behavior (as I will argue below in §2.3).

Endorsement also involves directing one's research on the basis of what is endorsed (condition 4). So, if a researcher endorses a theory, and there is a particularly important method for testing that theory, or a characteristic kind of methodology tied to the theory's tradition or heuristic, then this will affect what the researcher should be doing.[11] That is, the researcher should plan to engage in the method, or the associated methodology, because they endorse the theory. This, along with the strong obligation to defend, means that endorsement carries with it certain "pragmatic commitments" that belief does not.[12] Endorsing a theory means directing one's research toward that theory. Although beliefs clearly govern our behavior, believing that P does not generally commit one to researching about P.

The sixth condition concerns the fact that endorsement is an attitude one takes toward theories which are potential answers to the questions which guide a domain of inquiry. It is not the kind of attitude that an engineer takes toward Newtonian mechanics when using that theory for the purposes of bridge-building. For endorsement to play it's role in helping to motivate researchers, it must be that the researchers consider the theory in question to be a live option. A subject endorsing p advocates for p, and wants to see it turn out to be true. This is only possible if endorsements are limited to live options.

The seventh condition is about the aim of endorsement. The aim of endorsement is to promote healthy inquiry. "Healthy" here is just meant as a neutral term to express positive evaluation. Endorsement aims at the collective good of inquiry. Endorsing p aims at promoting the health of inquiry by allowing the subject to appropriately

---

[10] This obligation to defend is similar to the role Kaplan (1981a, 1981b) sees for his notion of acceptance. Later, however, Kaplan (1995) explicitly equates this kind of acceptance with belief, which I think is a conflation.

[11] For the notion of a heuristic, or characteristic methodology, see Whitt (1992) and Šešelja and Straßer (2014).

[12] For more on this notion of pragmatic commitment, see Van Fraassen (1980) and Whitt (1990).

engage in community discourse and debate, and by motivating her to commit to a research program which promotes the community's goals. This connects the aim of endorsement with truth, but only indirectly. Endorsing promotes the community's learning more truths.[13]

The last two conditions also help illustrate the way endorsement is distinct from belief. Belief aims at truth. It seems plausible that there is a sense in which believing falsehoods is simply impermissible, and it is this requirement that generates various other epistemic requirements. However, endorsement's connection with the truth (of the proposition endorsed) is more closely akin to guessing. When guessing, the hope is to guess the truth, but with little expectation of reliability in achieving this. It is permissible to guess (and endorse) unreliably. This is why endorsement is only appropriate to theories which are not known to be false, but these theories need not be the most likely theory to be true. In contrast, permissible belief requires some kind of reasonable expectation of achieving its aim of getting at the truth, and so believing unlikely things is inappropriate.

Another thing which distinguishes belief and endorsement is that there are different norms for rational belief than there are for rational endorsement. Where to draw the lines between what characterizes the attitude and what the norms of the attitude are is a bit difficult. There is significant overlap between the two projects. However, this does not seem terribly problematic, as belief is generally taken to have constitutive rationality requirements and a constitutive truth aim (Shah & Velleman 2005).

The biggest normative difference between endorsement and belief involves the following principle, which I take to be a bedrock intuitive assumption about belief: it is irrational to believe some proposition p if one takes $\neg p$ to be more probable than p. Put simply, you should not believe something you think is more likely false than true. This principle is not true for endorsement. One should not knowingly endorse something false, but one can endorse something unlikely to be true. Endorsement is an appropriate attitude for theories which should be pursued and advocated for, but

[13] I tend to think good inquiry is that which leads to the community learning interesting truths, but "healthy" here could refer to meeting a variety of epistemic standards.

which are (at least as yet) unconfirmed.[14]

With endorsement on the table, we now turn to the role it is meant to play in inquiry, which will in turn enable us to derive appropriate norms for rational endorsement.

There are two projects that recognizing endorsement contributes to:

Vindication Explain, rationalize, and justify our current practice.

Prescription Offer normative guidance to philosophers, humanists, and scientists about the appropriate attitudes to take to their favored theories.

There are both normative and descriptive elements to the vindication pro-ject: I want to offer a description that accurately reflects and explains the actions of at least some researchers, while also justifying them. However, the theory of endorsement also provides normative guidance for how to structure inquiry, and this is reflected in the prescription project.

These two projects give us some guidance on what an account of the norms of endorsement should look like. The norms need to be attainable by subjects, without the subjects' being specifically knowledgeable about "endorsement," which is not a univocal pre-theoretical notion. To pursue the vindication project, the norms need to be attainable, so they are met by some actual researchers. For the purposes of providing prescription and justification, the norms need to reflect the features of inquiry which we take to be valuable.

2.3 The Value of Endorsement

In this section, I want to briefly characterize the valuable features of inquiry that endorsement promotes.

[14] One might worry that this will permit endorsement of theories which should be ruled out, either for epistemic or moral reasons, e.g., that anthropogenic climate change is not occurring, or pseudo-scientific racist theories. However, there are two ways of resisting the idea that endorsing these theories would be appropriate. First, I think moral reasons are over-riding, so if one has a moral reason not to endorse a theory, this will mean that one should not do so, all-things-considered. Second, and more directly, endorsement is an attitude taken to live options, not options known to be false. The two examples here are both known to be false, and so are not potential candidates for endorsement. Thanks to Briana Toole for discussion on this worry.

Endorsement is governed by inclusive epistemic rationality, which includes extrinsic reasons. This is justified as a distinct and significant kind of rationality because it includes sensitivity to features of healthy inquiry. These are features which an accuracy-only notion of epistemic rationality leaves out. Because of this, endorsing in accordance with inclusive epistemic rationality promotes better inquiry. There are a number of such features that endorsement promotes.

A central feature of inquiry promoted by endorsement is the resilient commitment of researchers to their theories. It is undesirable for a researcher to drop a costly research program at the first sign of strong contrary evidence. We want researchers who are motivated to continue to defend their theories in the face of difficult objections, because sometimes objections and counter-evidence turn out to be misleading. Endorsement is an attitude that is characterized by this kind of commitment, and the norms of rational endorsement reflect this fact.

Having a resilient commitment to a theory means a researcher will commit significant time and energy to exploring implications of the theory. They will be motivated to develop the best version of the theory, the best possible defense of the theory, and to explain away anomalies and purported counter-examples.

Thus, having researchers who endorse their theories will contribute to the vivacity of debate. To borrow a famous notion, this will encourage a robust "marketplace of ideas" where views will get their full, fair hearing.[15] Researchers will be motivated to be strong advocates for their favored views, and allow them to have their "day in court." This kind of debate is valuable as a feature of inquiry, and a practice of rational endorsement is well-positioned to promote it. The norms of endorsement, therefore, should reflect this.

Similarly, it is valuable for inquiry to avoid certain kinds of undesirable deference on the part of researchers. Usually, when we encounter the advice of experts in a certain field (when we are not members of that field), we should defer to the experts'

[15]The analogy of free expression of ideas to an economic marketplace seems to trace back to Mill through Supreme Court Justice Oliver Wendell Holmes, though it is perhaps an imperfect metaphor for Mill's own view (Gordon 1997).

judgment. When a layperson discusses a theory's import to the broader society, they should defer to experts in the field. This also seems like a good rule for policy-makers. But it would be undesirable to have this deference when actually engaged in research. We do not want all researchers (even novices) to simply defer to the best experts in their field; this would stifle creativity and progress. This non-deference is clearly a necessary requirement of healthy philosophical discourse, for example. And I think the same is true throughout inquiry. Endorsement is also well-positioned to encourage this. Rational Endorsement involves advocacy of a theory, and may be justified even when one's confidence in the theory is lowered by disagreement.

It is highly intuitive that it is valuable to have researchers who serve as committed advocates for views, and who are willing to defend less well-developed views in the face of disagreement. In addition, there is significant empirical evidence that human inquiry and reasoning is improved by group arguments with these features. Groups that engage in debate, and begin with disagreement, are better at getting at the truth, and at producing instances of good argumentation. This has been supported by a number of psychological experiments involving a variety of reasoning and decision-making tasks (Bonner, Baumann, & Dalal 2002; Geil 1998; Kerr, MacCoun, & Kramer 1996; Kerr & Tindale 2004; Laughlin, Bonner, & Miner 2002; Laughlin & Ellis 1986; Mercier 2016; Mercier & Sperber 2011; Resnick, Salmon, Zeitz, Wathen, & Holowchak 1993). When people are challenged in their arguments (or even when they expect to be so challenged), they produce better arguments, recognize problems with others' arguments, and avoid making fallacious inferences.[16]

In addition to the empirical psychology literature, there is also support in the form of case studies from science. In particular, a recent paper by De Cruz and De Smedt (2013) appeals to a case study from paleoanthropology to argue that disagreement is valuable for inquiry. They consider the case of the alleged discovery of *Homo floresiensis*, a proposed hominin species. Paleoanthropologists disagree about whether the skeletal remains of small hominins found in a cave on the Indonesian island of Flores

[16]For an overview of the empirical literature regarding the benefits of debate and disagreement in group reasoning, see Mercier and Sperber (2011) and Mercier (2016).

are those of a new species, or whether they are actually the remains of pathological modern humans (i.e., humans with some sort of hereditary condition leading to small stature, myoencephaly, and various other slight differences from more typical human adults).

De Cruz and De Smedt argue that the disagreement among scientists over this question has led to three significant benefits. First, it results in the generation of new evidence, as proponents of each view are motivated to seek new and better evidence to convince their peers. Second, it leads to a reassessment of existing evidence and old assumptions. This is because it motivates scientists to look for ways in which old evidence might provide support for their view, and to look at old assumptions that conflict with their view, but are inadequately substantiated. Third, disagreement helps overcome confirmation bias, since researchers seek evidence and objections to views they oppose, and their disagreeing peers force them to notice and account for objections to their own view.

Endorsement enables the kind of resilient commitment and advocacy of a theory which leads to the valuable disagreement described by both the psychological literature and scientific cases studies.

An additional valuable features of inquiry will be explored in more depth below, in section 2.5: appropriate distribution of cognitive labor during the course of research.

2.4 Inclusive Epistemic Rationality

Ellie's case (from the very beginning of the paper) is an example of a familiar and widespread occurrence. There are many inquirers who are motivated to contribute positively to inquiry. They are not motivated by personal gain or by fame or fortune (or at least not *only* by such things). Rather, they pursue and commit to a theory partially out of a concern for contributing to healthy inquiry. Ellie just wants to do her part in getting at the truth. Reasons which bear on the health and success of inquiry are genuine epistemic reasons. A view which did not distinguish these reasons from

purely pragmatic ones would fail to make an important distinction.[17]

Endorsement is to be distinguished from belief, on the one hand, and from various practically-oriented acceptance notions, on the other, on the basis of the kind of epistemic rationality that governs it. What it is *epistemically* rational to endorse depends on reasons beyond those on which rational belief is based. Yet these reasons are also genuinely epistemic, as we see in Ellie's case. We can still distinguish between this kind of inclusive epistemic rationality and the "anything goes" pragmatist view of Rinard (2015).[18]

Rinard argues that there is no sense of epistemic rationality that can be distinguished from practical rationality (2015). Belief formation is governed by the same rational standards as any other act. Rationality is thus univocal: epistemic considerations are only relevant to an act's choice-worthiness to the degree that a subject happens to value them. I'm calling such a view "anything goes," because any reason or consideration relevant to the agent's interests gets counted in determining what is rational for that agent to do.

The anything goes view fails to adequately explain cases like Ellie's, and fails to distinguish them from cases where the subject is motivated entirely by pragmatic reasons. There is a clear sense in which Ellie is justified merely by appeal to considerations about what is good for inquiry. Of course, there might be pragmatic reasons for what she does as well (e.g., it might be good for her career). But there need not be any such practical reasons: Ellie would be justified merely by appeal to what is good for inquiry. Distinguishing between a sense of rationality which is inclusively epistemic, and a sense which is "anything goes," helps us to distinguish a case like Ellie's from one where the subject is motivated only by pragmatic concerns.

---

[17] Of course, none of this is to deny that some researchers really are motivated by prudential reasons, especially fame and prestige. This motivation is not even always a bad thing for science (see Kitcher (1990); Strevens (2003)). I am merely suggesting that some of us are sometimes motivated by a desire to contribute to inquiry.

[18] Throughout, I use terms like "reasons," "considerations," and "values," and treat them as though they are interchangeable. I think my view is compatible with a wide variety of views about the nature of normativity, so one can simply plug in one's favored view from the meta-ethics and meta-epistemology literature. For an overview of available theories, see Alvarez (2016); Broome (2015); Finlay and Schroeder (2015); FitzPatrick (2004); Gert (2009); Parfit and Broome (1997).

Contrary to the arguments of Rinard and other pragmatists, it has long been standard to suggest that there is a sense of epistemic rationality (or justification) which is importantly distinct from prudential rationality. That is, we can evaluate a belief based on purely epistemic merits. However, what considerations, values, and reasons count as epistemic, in this sense, is plausibly more narrow than the inclusive epistemic rationality I will propose below.

Examples of this kind of epistemic consideration are how well a belief is supported by evidence, how coherent a set of beliefs is, or how reliably they were produced. Such concerns might be overridden by moral and even sometimes prudential considerations, but they are clearly separable from non-epistemic considerations.[19]

When evaluating beliefs (and degrees of belief) epistemically, it is plausible that the only relevant considerations concern accuracy. That is, the rationality of a belief is determined solely by features that reflect on *its* truth, or how likely it is to be true. Similarly, whether one's credences are rational is determined by their accuracy (i.e., by how close they are to the vindicated credence function) and features which indicate accuracy. There are two primary motivations for thinking this. The first is the recent success of accuracy-based epistemic utility theory. This program has generated arguments for probabilism, conditionalization, and various other coherence norms by appeal to accuracy alone (Easwaran 2013, 2016; Fitelson 2016; Fitelson & Easwaran 2015; Greaves & Wallace 2006; Joyce 1998; Konek & Levinstein 2016; Pettigrew 2016).

The second motivation for thinking that the epistemic rationality of belief and credence is limited to accuracy involves epistemic bribery counter-examples to epistemic consequentialism for belief (Berker 2013; Firth 1981; Greaves 2013; Jenkins 2007). These are cases of intuitively impermissible epistemic bribery designed to show that it is irrational to believe something on the basis of epistemic gains later, or elsewhere in one's belief state. That is, it is irrational to trade having a false belief now to gain more truths later, or to trade a false belief about P for true beliefs about Q and R. In Firth's original example, he describes an atheist researcher who has the opportunity

[19]For a small sample of arguments in favor of a distinctively epistemic domain of evaluation, see Goldman (1986); Shah and Velleman (2005); Sosa (2009, 2015).

to gain a great many further truths by taking on a belief in god (which he thinks is false).[20]

Intuitively, the atheist scientist is unjustified or irrational in believing in God in order to gain further epistemic value down the line. There is something inappropriate about accepting epistemic badness now for epistemic goodies later. This means that not only is the epistemic rationality of a belief sensitive only to accuracy considerations, the rationality of any particular belief is only sensitive to considerations concerning the accuracy of *that very belief* (and *mutatis mutandis* for degree of belief).

Endorsement, on the other hand, should be sensitive to considerations beyond the truth of the particular proposition in question. As an attitude meant to facilitate inquiry, it needs to be sensitive to the valuable features of inquiry described above (in section 2.3). In contrast to belief, it can be rational to endorse a proposition because it will lead to downstream epistemic goods, even where that comes at the potential cost of the accuracy of the particular attitude. So, it might be rational for the atheist scientist to *endorse* theism, even in the face of the evidence they take themselves to have for atheism, precisely because this will lead to the improvement of inquiry in general. So while I agree about the irrationality of *belief* in the atheist scientist's case, I think there is an attitude to theism that it is appropriate for the scientist to take in order to learn more in the future: endorsement.

Thus, we need a kind of rationality that is sensitive to considerations besides accuracy, but remains purely epistemic. I call this *inclusive epistemic rationality*. We can characterize this notion by appealing to a distinction between *intrinsic* and *extrinsic* epistemic reasons (or values), inspired by Daniel Steel (2010).[21] An epistemic value

---

[20] Borrowing the formulation in Berker (2013):

"Suppose I am a scientist seeking to get a grant from a religious organization. Suppose, also, that I am an atheist: I have thought long and hard about whether God exists and have eventually come to the conclusion that He does not. However, I realize that my only chance of receiving funding from the organization is to believe in the existence of God: they only give grants to believers, and I know I am such a bad liar that I won't be able to convince the organization's review board that I believe God exists unless I genuinely do. Finally, I know that, were I to receive the grant, I would use it to further my research, which would allow me to form a large number of new true beliefs and to revise a large number of previously held false beliefs about a variety of matters of great intellectual significance. Given these circumstances, should I form a belief that God exists? Would such a belief be epistemically rational, or reasonable, or justified?"

[21] Steel's discussion of this idea is brief, so I am uncertain whether my use of this distinction precisely

is *intrinsic* if manifesting the value constitutes an attainment of truth, is necessary for truth, or indicates truth. A reason is intrinsic if it is a reason to believe that a proposition is true, or a reason to accept it as true. Being true is an intrinsic epistemic value, as is consistency.[22]

Epistemic reasons are extrinsic when they are about features which tend to promote attaining the truth, but are neither necessary features of truth, nor reliable indicators. An example Steel suggests is testability (2010, 15). That a hypothesis or theory is amenable to testing is not a reliable indicator that it is true; many testable claims are false. Instead, valuing testability is a methodological commitment which promotes the truth. Considerations of what makes for valuable, productive inquiry are extrinsic reasons. That there is vivacity of debate, or that researchers are motivated to defend a theory, are not necessary conditions of truth. Rather, they promote truth in the long run. In order for endorsement to play the appropriate role in inquiry that I suggest it does, it needs to be sensitive to such considerations. It needs to be sensitive to extrinsic epistemic reasons, as well as intrinsic ones. Hence, it is governed by an inclusive epistemic rationality: one that includes both intrinsic and extrinsic reasons.

Extrinsic epistemic reasons are reasons for endorsing a theory which stem from the fact that doing so will promote healthy inquiry. They are reasons which (indirectly) promote epistemic goals. However, this is not yet enough to distinguish them from pragmatic reasons. In order to more fully characterize and distinguish this kind of reason, I will provide a response to a potential objection to the idea that these reasons are genuinely epistemic.

This objection concerns the way extrinsic epistemic reasons promote epistemic

---

tracks his (2010, 18). Jenkins (2007) appeals to a distinction that is very similar to mine. She distinguishes between "extraneous consequences" of a belief in P, and those which "which directly concern P itself" (37).

[22]The notion of intrinsic epistemic value is related to the idea that belief is "transparent," in the sense explored by Shah and Velleman (2005). Transparency here means that the reasons to believe p are simply reasons for p, or evidence for p. Whenever one considers the question of "whether to believe p," this question is equivalent to the question of "whether p." Beliefs are transparent is because they are only appropriately sensitive to intrinsic epistemic values.

goals of inquiry. One might admit that these are genuine reasons, because they promote good epistemic outcomes. But then one might still deny that they count as epistemic, because of the way this promotion works. Instead, it might be suggested that these are really pragmatic reasons; just ones aimed at epistemic ends. This worry can be illustrated by appeal to cases which involve "sandwich reasons."[23]

Suppose (plausibly enough) that hunger degrades intellectual performance. Then, eating a sandwich will count as promoting healthy inquiry, because it will lead to better intellectual performance. Suppose also that, if I endorse theory A, this will involve working in a lab near a good sandwich shop, so that I will be less hungry and thus (slightly) better in intellectual performance each day at work.

The worry is that my theory of inclusive epistemic rationality seems to predict that I have a sandwich-based (extrinsic) epistemic reason to endorse theory A, but intuitively this is not an epistemic reason. That is, it is intuitively implausible that I could have a reason to endorse a theory based on the fact that doing so will get me more sandwiches (even if the sandwiches really will improve my epistemic performance).

Happily, there are additional resources to draw upon in distinguishing extrinsic epistemic reasons from mere "sandwich reasons." Extrinsic epistemic reasons are reasons which are internal to a domain of inquiry. Whether something counts as a reason within a certain domain depends on both the goals of that domain and the internal standards of that domain. Performances and states within a domain are evaluated based on whether they promote the goal of the domain, and on whether the way this goal is promoted meets the standards of the domain.[24]

This way of categorizing reasons can be illustrated by appeal to the domain of chess. The (proximate) aim or goal of a player in chess is winning the game. In order to be a chess-reason, something must promote this goal. For instance, if castling, given the circumstances, will increase the likelihood of my winning the game, this is a chess-reason for me to castle. If I decide to castle because I know it will impress my friend

[23]I learned of cases like this one from Nomy Arpaly (2017), who attributes them to Sophie Horowitz. The two objections I am considering here are largely inspired by Arpaly's paper.

[24]Here I am drawing from Sosa (2015), especially Chapter 8.

who is watching the game, this is not a chess reason because it does not promote the aim of winning.

There are, however, ways of failing to be a chess reason other than not aiming at the goal proper to the domain. Suppose I have a tattoo on my arm, and if I move my Rook to my opponent's side of the board, this will reveal the tattoo to him. I know that this will intimidate him and cause him to play cautiously and predictably. Thus, the move will promote the goal of winning the game. However, it does so in a way which violates the internal standards of the practice of chess. Intimidating an opponent is not a move which is proper within the domain of chess. If I were playing a different game, this might be totally acceptable (e.g., in the board game *Diplomacy*, which notoriously involves manipulating and deceiving one's opponents). But the standards internal to chess rule this out as a chess-appropriate move.

Domains of inquiry are like chess: they have their own standards which help determine what counts as a reason. Precisely what the internal standards are in research domains is a complicated question, and one I don't think we can answer in full generality. Different domains have different standards based on the kind of research involved. This fact is familiar: it is well-known that different fields have varying standards of evidence. Thus, what counts as an *intrinsic* epistemic reason is clearly determined in part by internal standards of the field of inquiry (though in a way that cannot be totally disconnected from truth if the field is well-functioning). I am suggesting that the same is true for extrinsic epistemic reasons.

Precisely how the internal standards of a research domain are set is a question I will remain neutral on. However, there are at least two clear options. First, the standards could be set merely via convention, or social construction: the social practices of the researchers determine what the standards are.[25] The other way the standards could be set is by appeal to the essential features of a particular kind of research. The thought here is that research fields involve investigation of different kinds of phenomena, and investigation using different methods, in a way that is non-arbitrary. So the

[25] I think Sosa (2015) has something like this in mind for determining epistemic standards for belief-formation performances.

standards of astrophysics will be different than the standards of biochemistry, for reasons having to do with the nature of the phenomena investigated. The essential nature of these investigations then determines the appropriate internal standards of the field. Either way of standard-setting (socially constructed or essential) will work to rule out sandwich reasons.

Domain internal standards will rule in, as extrinsic epistemic reasons, considerations like distribution of resources (and of cognitive labor), but rule out "sandwich reasons." This is clear from scientific practice, and explains our intuitions about sandwiches. Choosing a theory on the basis of the proximity of the relevant lab to a sandwich shop will fail to meet the internal standards of a scientific research domain.

In light of the responses to this objection, we have the following characterization:

> **Extrinsic epistemic reasons** are reasons which indirectly promote epistemic goals of inquiry, and promote them in a way that meets the internal standards of the relevant domain of inquiry.

The expanded, inclusive epistemic rationality of endorsement is part of what makes it a more appropriate attitude to take during the course of inquiry. As we will see, it will allow us to appeal to resources unavailable for belief, including a type of epistemic utility and decision theory that is not appropriate for belief.[26]

Below, I will present a few particular examples of extrinsic epistemic reasons, and show how we can build a decision theory for rational endorsement which incorporates appropriate sensitivity to extrinsic epistemic considerations. In particular, I will focus on some extrinsic reasons provided by insights in the social epistemology literature. The appeal to endorsement, as well as the appeal to inclusive epistemic rationality, is justified in part because it includes additional considerations that are clearly relevant to our epistemic practice, but which are left out of a traditional, belief-based view.[27]

---

[26] In order to side-step worries about attitude voluntarism , we can treat inclusive epistemic rationality as providing evaluative standards (rather than deontic norms). This is a common move to make in epistemology: see, e.g., Fitelson and Easwaran (2015). For more on the distinction between deontic and evaluative norms, see Smith (2005).

[27] Although there is not room to explore the thought here, I think it is (at least very often) *irrational* to believe one's favored theory. In brief, there are three main considerations that should lower

2.5 Extrinsic Reasons from Social Epistemology

Endorsement is the appropriate attitude to have toward favored theories during the course of inquiry. This is due in part to its sensitivity to extrinsic epistemic reasons. This sensitivity allows us to solve several epistemological problems in ways that are not available to accounts that only recognize belief. Specifically, it eases certain tensions between individual rationality and collective rationality in inquiry.

One social epistemology issue that endorsement can help with is the appropriate distribution of cognitive labor. The contemporary literature on the topic began with Kitcher's (1990) paper. One of the primary concerns of this literature is an apparent tension between individual and collective epistemic rationality. Once we recognize endorsement, and its sensitivity to extrinsic epistemic reasons, this apparent tension can be easily resolved.

The problem is this: it seems that the rational thing to do, epistemically, is to pursue the most probable theory. However, if every researcher individually follows this advice, it will lead to clearly bad distributions of labor. To see this, suppose that we have two candidate theories, A and B. A is 80% likely to be true while B is 20% likely. We do not want every researcher working on A; after all, A might be false. It would be better to have at least some researchers working on B. But if every researcher does the (apparently) individually rational thing, then all researchers will work on A, and none on B.

In order to resolve this apparent tension, Kitcher (1990) and Strevens (2003) appeal to economic modeling, and the priority rule in science (the convention of awarding all credit for a discovery to the individuals or lab that first succeeds in discovering it). First, they use an economic model to find the optimal distribution of labor, given certain assumptions. Then, they argue that the prestige-seeking behavior of scientists,

---

our confidence in theories in cutting-edge research domains: pervasive disagreement, the pessimistic meta-induction, and under-determination of theory by evidence. These problems are characteristic of cutting-edge domains, and so the subjective probability (confidence, or credence) we assign to the theories should be too low to justify belief. Indeed, in many such cases our confidence in the theory should be less than half, in which case full belief is clearly unwarranted. Moreover, as I have argued, the epistemic rationality governing belief is not sensitive to extrinsic reasons which do and should govern our decisions about which theories to be committed to, and to pursue.

coupled with the priority rule, will lead to (or at least promote) this distribution of labor.

We can call this kind of economic model of science a *marginal contribution reward* (MCR) system (following Muldoon (2013)). This method considers rules of behavior as they apply to representative members of a research community. The goal is to determine what sort of reward structure is necessary to make it individually rational for a subject to behave so that, when combined with everyone else's, the behavior contributes to the optimal distribution of labor. If the reward structure is appropriately constructed, then it will be rational for an arbitrary, representative member of the research community to behave in a way conducive to community goals.

For ease of exposition, I will focus on a particular version of the MCR strategy, the one developed by Strevens (2003).[28] This is not because I am endorsing Strevens' approach (though I admit some affinity for it). Rather, I am appealing to it as an example, to show that endorsement is compatible with solutions to the distribution of cognitive labor problem, in a way that belief cannot be. I am offering a "proof of concept." I want to show how to implement solutions from the social epistemology literature in the account of individual epistemic rationality using endorsement.

Strevens, like Kitcher, appeals to pragmatic considerations in his MCR account (Kitcher 1990; Strevens 2003). Specifically, he suggests that researchers should be (and are) rewarded with prestige in proportion to their contribution to a project. Without this appeal to pragmatic considerations, MCR accounts cannot deliver the goods of individually rational behavior leading to collectively rational results. In order to achieve collective *epistemic* rationality, they appeal to individual *practical* rationality. However, by appealing to endorsement, we can instead represent the subject's individual rationality as taking the collectively rational into account, and we can do this by appealing to the same conceptual resources as already used by the MCR account.

In short, what I will show is that we can build the reward structure recommended

[28] I focus here on MCR models, however, other kinds of modeling might also be useful for discovering extrinsic epistemic reasons, e.g., Agent-Based epistemic terrain modeling (Muldoon 2013; Muldoon & Weisberg 2011; Thoma 2015; M. Weisberg & Muldoon 2009; Zollman 2009). The endorsement framework could easily implement constraints derived from such modeling approaches.

by the MCR account into the epistemic utility function of the researcher. This is possible because endorsement is appropriately sensitive to extrinsic epistemic values.

In pursuit of this project, I will briefly describe Strevens' version of MCR. The decision-theory for endorsement incorporating MCR will be explored in section 2.6.

Suppose there are two scientific research projects, P_1 and P_2. There is a "success" function for each project, describing how likely the project is to produce truth given a number of people working on it. That is, it takes a number of researchers (n) as input and outputs the probability that the project will be successful.[29] Other things being equal (it seems plausible to assume), projects that are more likely to bear truth will have bigger values for any particular number of researchers n. Call the success functions for P_1 and P_2, $s_1(\cdot)$ and $s_2(\cdot)$, respectively. Suppose also that P_1 is the more likely project to pay off, and so $s_1(n) > s_2(n)$. As mentioned above, there will be an optimal distribution of researchers based on these success functions: where P_1 is more likely to get us the truth, and N is the total number of researchers (or research hours), what is to be optimized is:

$$s_1(n) + s_2(N - n)$$

This will give us the best chance of getting at the truth.

The problem, as discussed above, is that on their own, if each researcher is trying to give themselves the best opportunity of personally getting to the truth, they will all choose to work on the project with the better success function: the one more likely to get at the truth. In this case, this will mean everyone is working on P_1, and this will not (generally) be the optimal distribution. Inquiry as a whole is not best served by having all researchers working on the same project.

Strevens shows that what we need in order to reach the optimal distribution is for each researcher to maximize their own *marginal contribution*. Here, "marginal contribution" means the increase to the probability the project will pay off that is provided by the researchers joining the project. That is to say, my marginal contribution is how much more likely the project is to pay off after I work on it. It is the difference I

[29] This can also be done in terms of research hours, rather than individuals. Number of researchers is more natural for my account, but either will work.

make to the probability of success. We can represent marginal contribution of P_1 as a function $m_1(\cdot)$, defined as:

$$m_1(n) = s_1(n+1) - s_1(n)$$

Strevens shows that rewarding the researcher based on their marginal contribution, a reward scheme he calls *Marge*, will lead to a better distribution of labor. This claim is based on a few plausible assumptions. In particular, it requires that the success functions are increasing, but have decreasing marginal gains. That is, every researcher added to the project increases its probability of paying off. However, each additional researcher adds less probability than the last one, i.e., the project becomes saturated. This seems like a perfectly reasonable assumption: the more researchers already working on something, the less good it will do to add another one.[30]

So in our example, when more researchers are already working on P_1, the marginal contribution for a new worker joining it will be smaller. Meanwhile, the marginal contribution for joining P_2 will be higher. That is, as n, the number of researchers working on P_1 increases, $m_1(n)$ decreases and $m_2(n)$ increases. This can result in its being much more lucrative to work on P_2, even though the overall probability of its paying off might be significantly lower. Thus, implementing the *Marge* reward scheme will lead to better distributions of labor, since researchers will be incentivized to work on projects with less probability of paying off, but to which their own work will contribute much more.[31]

One benefit of Strevens' theory is that it vindicates aspects of scientific practice. Given human nature, it seems good that we can implement practical rewards that make individual practical rationality line up with the collective good (a notion familiar from economic theory). However, prudential motivations can lead to distorting influences, as researchers have monetary and career incentives to *seem* to contribute

[30] These assumptions can be relaxed to obtain very similar results (Kitcher 1990).

[31] Kitcher and Strevens appeal to an MCR scheme in order to vindicate the priority rule in science. I will leave aside the differences between the *Priority* reward scheme and *Marge*, as the details do not affect the project here.

to inquiry without actually doing so (fraud, bribery, plagiarism, etc.).[32] This will result in some failures of Strevens' idealizing assumptions. So we might need some additional work to ensure collective rationality, beyond simply appealing to practical considerations.

More importantly, however, researchers motivated by practical concerns should not be the only ones who can rationally act so as to benefit inquiry as a whole. Being motivated by personal gain should not be a requirement of *epistemic* rationality! Instead, we should borrow the insights of the MCR tradition, in order to make individual *epistemic* rationality consonant with collective epistemic rationality. We should be able to characterize the purely epistemic norms in such a way that someone with only epistemic motivations can, also, *rationally* contribute to a healthy distribution of cognitive labor. Thanks to inclusive epistemic rationality, with its sensitivity to extrinsic reasons, we are in a position to do so.[33] This is because facts about what will lead to a better distribution of labor are extrinsic reasons which affect what it is rational to endorse. In order to reflect this, below I offer a decision-theoretic account of rational endorsement which builds a sensitivity to marginal contribution into the epistemic utility function for researchers. This provides a formal account of individual epistemic rationality which coheres with collective epistemic rationality.

2.6 Rational Endorsement

Rational endorsement is governed by inclusive epistemic rationality. The picture I have been developing comes to this: the rational theory to endorse is the one supported by the weight of both intrinsic and extrinsic epistemic reasons. How should we weigh these reasons when making a decision about what to endorse? One plausible answer is to turn to our standard theory of weighing considerations for action: decision theory.

---

[32] Though see Bright (2016) for how "pure" alethic goals can lead to fraud, too.

[33] It is worth noting that Kitcher briefly mentioned a solution somewhat similar to mine, but dismisses it as "redefinition" (1990, 14). I think this is a mistake. The project is not merely to stipulate that individual rationality is sensitive to concerns of collective rationality, but to show the explanatory payoff of a theory which ties them together in a coherent and rigorous manner.

In this section, I will give a formal characterization of rational endorsement, using the tools of decision theory. This account is meant to reflect the norms and features of endorsement that have already been characterized. The basic idea is that one should endorse a theory just when doing so will maximize expected inclusive epistemic utility.

For clarity and ease of exposition, I am going to first provide an explanation of how the decision theory works (§2.6.1). I will then give the formal characterization in a separate subsection (§2.6.2), for readers who are interested in the technical details.

2.6.1 Rational Choice of Endorsement

The decision theory for endorsement is an epistemic one, because it involves an epistemic utility function. However, the epistemic utility here is distinct from the kind currently most popular in the literature. The contemporary notion of epistemic utility is "accuracy-first." On this view, accuracy is the sole determinant of how epistemically valuable an outcome is. Sets of beliefs, or credences, have a certain value based on how close they are to the truth. This kind of epistemic utility theory was pioneered by Joyce (1998), and has been fruitfully applied in a large, rapidly expanding literature.[34]

Accuracy-focused utility theory is perfectly appropriate for determining norms of belief and credence, because such attitudes are appropriately sensitive only to intrinsic epistemic reasons. However, since rational endorsement is governed by inclusive epistemic rationality, we need a more inclusive notion of epistemic utility. Helpfully, there is an older tradition of epistemic utility that we can appeal to, owing especially to the work of Levi (1974, 1980, 2004a), Kaplan (1981a) and Maher (1993). The theory I am giving here is largely inspired by Maher's.

Using Maher-style epistemic utility, we begin with a standard subjective utility function for an agent, one that reflects the agent's values and desires. We then add constraints on the utility function which ensure that it is appropriately epistemic.

[34]For more on this kind of epistemic utility, see Easwaran (2013, 2016); Fitelson (2016); Fitelson and Easwaran (2015); Greaves and Wallace (2006); Joyce (1998); Konek and Levinstein (2016); Pettigrew (2016).

These constraints require that the subject values truths over falsehoods, values the avoidance of contradictions, and values interesting truths over prosaic ones. This was Maher's strategy for ensuring that the subject would have "scientific values." What I want to add to this strategy is additional constraints which will ensure that the subject is appropriately sensitive to extrinsic epistemic values. The new constraints that I add below are designed to implement the solutions borrowed from the social epistemology literature reviewed above, and in general to make the agent sensitive to both intrinsic and extrinsic considerations.

The version of the decision theory I offer below employs only a small number of constraints. It includes three of Maher's constraints. The first of these is a weak truth constraint: that the subject prefers to endorse a theory a when a is true. This constraint is meant to help ensure that researchers only endorse theories which are live options: theories that have some probability of turning out to be true, even if they are less likely than their negation, or than their competitors. This constraint is weak, however. It is compatible with a greater utility for endorsing b over a, even if b is false and a true. This weakness in the constraint allows the decision theory to model the fact that it can be rational to endorse something even when it is unlikely. This fits with endorsement's aim at promoting healthy inquiry.

The second constraint encodes a preference for information: a preference for more informative theories over less informative ones, as long as the more informative theory is not contradictory (since contradictions are guaranteed to count as highly informative, since they entail everything). The third constraint adds a general prohibition on contradiction.[35] These constraints, along with the standard norms of rational credence, are meant to capture the sensitivity to intrinsic epistemic reasons.

---

[35]I think this constraint will be operative in most domains. However, there are a number of specific research domains where this might need to be relaxed. For instance, we might need to relax it in characterizing the early stages of the pursuit of quantum mechanics, where the initial theory was known to be inconsistent (Faye 2014). More obviously, research about the applicability of paraconsistent logic, and dialetheism will violate this constraint (Priest 2006; Priest & Berto 2013; Priest, Tanaka, & Weber 2015). Even if all inconsistent theories are in fact false (because inconsistent), we can still model them using inclusively rational endorsement, as long as we relax this constraint. Thanks to Eddy Chen, Graham Priest, and Branden Fitelson for discussion on these points.

In addition to the constraints from Maher, there are two constraints meant to enforce sensitivity to extrinsic epistemic considerations. The first is taken from Strevens' *Marge* reward scheme, which I will simply call the *MCR constraint*. It ensures that the subject prefers to have a higher marginal contribution to their project. That is, they gain more utility from working on a project to which their own work will contribute more. This constraint appeals to the notion of a marginal contribution, and to a success function, described above (§2.5). To satisfy this constraint, the subject's utility will be higher when their marginal contribution is higher.

This constraint thus builds Strevens' solution to the distribution of cognitive labor directly into the account of individual rationality. What it is rational to endorse depends on one's marginal contribution. This is justified because the epistemically-motivated rational inquirer values being part of healthy inquiry, and given Strevens' arguments, striving to contribute the most will lead to better distribution of labor. Endorsement must be sensitive to these considerations, since it is associated with research planning and pursuit judgments.[36]

The second new constraint is meant to reflect the resilient commitment of endorsement. This constraint, which I call *Inertia*, requires that the subject's utility for endorsing a proposition significantly increase after they endorse it. Then, in any subsequent decisions about what to endorse, the previously endorsed proposition's utility will outweigh even significant drops in the probability of that proposition. Thus, the subject's endorsement will survive even very difficult evidence against the theory. So the resilient commitment which characterizes endorsement will have the appropriate rational standing.

I suspect these constraints will prove adequate to capture all of the reasons, both intrinsic and extrinsic, relevant to endorsement.[37] This is because the MCR constraint makes the subject sensitive to a variety of other extrinsic epistemic reasons. Anything

[36] As I suggest above, Strevens' model might not turn out to be the best one. If so, we can simply adopt the better model and give it the same treatment. Again, the purpose here is not to come down in favor of one solution in that domain, but to show how we can use such solutions in a framework for rational endorsement that smooths the tension between individual and collective epistemic rationality.

[37] Assuming, as above, that Strevens' MCR is the right view of how to ensure appropriate distribution of labor.

that makes the researcher more productive or effective will alter the success function, and so alter the researcher's marginal contribution. This means that extrinsic reasons will impact the success function of theories, as well as the degree of marginal contribution of that researcher in particular. So, sensitivity to marginal contribution thereby involves sensitivity to a variety of extrinsic reasons which impact the success function of the theory.

For example, extrinsic reasons such as the local availability of resources, or the particular talents of the researcher in question, will affect the researcher's own marginal contribution. Whether a researcher has access to an fMRI machine might genuinely impact how valuable for inquiry it would be for that researcher to pursue a project involving a theory that is best tested by fMRI experiments. And a researcher sensitive to their own marginal contribution will avoid endorsing a theory, and planning a research program, if they do not have access to the right equipment.

Similarly, a researcher's available network of contacts with other researchers, in her own field and relevant other fields, will affect how much marginal contribution she makes to the project of confirming her endorsement. Even just how excited a researcher is about a particular theory, or the kind of work necessary to test it, can make a relevant impact. People are more motivated to work on things they like, so their own affinities are legitimate extrinsic epistemic factors.

Thus, although these constraints were designed to allow the account of endorsement to help with distribution of cognitive labor, it also will help with ensuring sensitivity to other kinds of extrinsic epistemic reasons.

Though I prefer the epistemic utility function utilizing just these constraints, the framework developed is actually more flexible than this. It can accommodate treating other considerations as independent constraints on the utility function.[38]

The purpose of the decision theory for endorsement is not so that scientists can sit down and calculate what it is most rational to do. Rather, it serves as a set of evaluative constraints on researchers' preferences between different choices of endorsements. I

[38]For details on how this would work, see footnote 42.

think that one of its more valuable applications is to allow us to evaluate the actual decision-making procedures that researchers use in theory choice. This is much the way that the heuristics and biases literature in social psychology uses standard decision theory as its theory of rationality, and then uses this to evaluate actual human decision-making, and determine where such decision-making falls off the rails.[39]

2.6.2 Decision Theory for Endorsement

Here, I provide the formal details of the epistemic decision theory. The underlying formal framework for this is borrowed from Joyce (1999). For ease of exposition, this theory is evidential.[40]

In the decision theory for endorsement, each decision problem, $\mathbf{D} = (\Omega, \mathbf{A}, \mathbf{S}, \mathbf{O})$, is composed of these elements:

1. A partition of acts of endorsement, $\mathbf{A} = \{A_a, A_b, A_c, ...\}$, where A_a is the act of endorsing the proposition a.
2. A partition of states of the world, $\mathbf{S} = \{S_1, S_2, ...S_n\}$.
3. A partition of outcomes, $\mathbf{O} = \{A_i \& S_i | \forall A_i \in \mathbf{A} \; and \; \forall S_i \in \mathbf{S}\}$.
4. A σ-algebra Ω such that it is the smallest such algebra containing **A**, **S**, and **O** and their closure under negation and disjunction.
5. A credence function $P(\cdot)$ defined on Ω.
6. A utility function defined over **O** so that the utility of endorsing a in state S_1 is $u(A_a \& S_1)$.

The decision rule is then the standard requirement to choose the act that maximizes expected value:

$$EV(A_a) = \sum_{S_i \in \mathbf{S}} P(S_i|A_a) u(A_a \& S_i)$$

---

[39] For an overview this literature, see Kahneman (2013). The literature began with Tversky and Kahneman (1975). For a philosophical application of this idea, see J. Weisberg (2016).

[40] Expanding this to a causal decision theory is a relatively simple matter, but it adds some complications to the formalism which are irrelevant to our purposes here. For the procedure for the expansion to CDT, see Joyce (1999).

The act-space here consists of acts of forming endorsements, indexed by theories or propositions that count as eligible options for endorsement. This will be domain-specific: in each research domain, there will be considerations appropriate to that domain that researchers will need to be sensitive to in deciding what the available options are.

My theory introduces no new difficulties with the probability function and state-space, and we can treat these in the standard way. The state-space can be the set of possible worlds (R. Stalnaker 1999; R. C. Stalnaker 1987), or some other partition that appropriately accounts for the researchers uncertainties (Jeffrey 1990; Joyce 1999).[41]

What sets this theory apart from a standard evidential decision theory is the constraints on the utility function. I will call them "M-constraints" after Maher (1993). They ensure that the theory will be appropriately sensitive to epistemic concerns, both intrinsic and extrinsic.

The set of M-Constraints:

1. **Respect for Truth**: $u(A_a \& S_i) \geq u(A_a \& S_j)$ for all $S_i \in a$ and all $S_j \notin a$ (i.e., higher utility whenever the state is one in which a is true).

2. **Respect for Information**: $u(A_{a\&b} \& S_i) \geq u(A_a \& S_i)$, for all $a, b \in \Omega$ s.t. $a \& b \nvDash \varnothing$. (i.e., prefer more specific, informative things, as long as they are not contradictory).

3. **Contradiction Suboptimal**: $u(A_a \& S_i) \geq u(A_\varnothing \& S_i)$ for any S_i and any a.

4. **MCR**: the utility of a state where the subject's marginal contribution is higher is at least as great as one where it is lower. Moreover, the difference in utility between two states should increase proportionately with differences between the marginal contribution of the agent in those states. Expressed formally:

   (a) When $S_i, S_j \in a$, and $m_a(A_a) = I$ in S_i, and $m_a(A_a) = J$ in S_j, and $I > J$, then

[41] The current version of the theory, since it is an evidential decision theory, makes this easier in some respects because it is partition invariant.

$u(A_a\&S_i) > u(A_a\&S_j)$ and

$$\frac{u(A_a\&S_i)}{u(A_a\&S_j)} \propto \frac{I}{J}$$

(b) When $S_i \notin a$ and $S_i \notin b$, if $m_a(A_a) > m_b(A_b)$, then $u(A_a\&S_i) > u(A_b\&S_i)$ and

$$\frac{u(A_a\&S_i)}{u(A_b\&S_i)} \propto \frac{m_a(A_a)}{m_b(A_b)}$$

5. **Inertia**: The utility of endorsing a theory becomes much higher once one endorses it:

   Where u_1 is utility function at time t_1: If the subject comes to endorse a between t_1 and t_2, then $u_2(A_a\&S_i) \gg u_1(A_a\&S_i)$, for all S_i.

With these constraints, we have an inclusive epistemic utility function. This provides a helpful characterization of rational endorsement.[42]

2.7 Conclusion

A field of inquiry in which researchers endorse their theories will be a healthier, more productive field. Recognizing endorsement, along with its associated normative framework, allows us to complete both the vindication and justification projects introduced in section 2.3. We can justify the actions of excellent researchers like Ellie, who are resiliently committed to their own theories, and who advocate strongly for them, even though they should not believe them. It also provides normative guidance for researchers who want to contribute to healthy, productive fields of inquiry.

---

[42] Although I prefer a theory which uses just the few constraints listed above, the framework is actually more flexible than this. There is a simple way to expand the decision theory to take into account additional extrinsic reasons more directly. We can do this by following Levi (1974, 1980) and Pettigrew (2014), and using a composite utility function. This function is composed of the weighted average of several sub-utility functions, each of which represents sensitivity to a different extrinsic reason.

Suppose $u_{mcr}(\cdot)$ is the sub-utility function structured as in the above. Also, $u_{heu}(\cdot)$ is the sub-function structured by heuristic power, where $u_{heu}(A_a\&S_i) > u_{heu}(A_b\&S_i)$ just when the heuristic of a is better (or more powerful) than b's. Furthermore, let $u_{exp}(\cdot)$ be the sub function structured by the novelty of explanations, where $u_{exp}(A_a\&S_i) > u_{exp}(A_b\&S_i)$ just when a provides more novel explanations of phenomena than b.

To obtain the overall epistemic utility function, $U(\cdot)$, we weight these individual sub-functions, then add them together. Let $\alpha_1 + \alpha_2 + \alpha_3 = 1$, where the size of each α is determined by how important the subject takes the different considerations to be. Then,

$$U(A_a\&S_i) = \alpha_1 u_{mcr}(A_a\&S_i) + \alpha_2 u_{heu}(A_a\&S_i) + \alpha_3 u_{exp}(A_a\&S_i)$$

Finally, this composite utility can be plugged into the expected utility calculation, as in §2.6.2.

The theory of rational endorsement provided here also shows how endorsement can help smooth the tension between individual and collective rationality. It provides an avenue for doing so that would not be available to a theory relying instead on belief. It achieves this by appeal to inclusive epistemic rationality, which is sensitive to both extrinsic and intrinsic epistemic considerations. This sensitivity is modeled by an epistemic decision theory which uses a set of constraints on the utility function.

Chapter 3

How to Endorse Conciliationism

3.1 Introduction

The epistemology of disagreement is concerned with the question of how you should respond, by the lights of epistemic rationality, to the discovery that you disagree with another person. This question is not about how you should respond to arguments or evidence offered by your interlocutor, but rather about how you should respond to the fact of the disagreement itself. The general view which suggests that (*ceteris paribus*) you should change your beliefs, merely on the basis of the discovery of the disagreement itself, is called *conciliationism*. Conciliationism, in both its more extreme and more moderate forms, has significant intuitive plausibility and is widely popular. However, it is beset by several difficult problems. In particular, it suffers from what I will call the *skeptical results problem* and the *self-undermining problem*.

In this paper, I will argue that we can solve both problems for conciliationism by appealing to the theory of *endorsement*. Endorsement is a propositional attitude that is distinct from both categorical (or full) belief and from degrees of belief, and which is governed by an inclusive epistemic rationality.

Below, I will first offer some background on conciliationism, and then describe the two problems in greater detail. Then, I will offer a characterization of endorsement and the inclusive epistemic rationality which governs it. With this background in place, I will show how the theory of endorsement solves each of the two problems for conciliationism.

3.2 Conciliationism

Consider the following version of a classic case from the epistemology of disagreement (Christensen 2007):

> **Mental Math:** Suppose you go to lunch with your friend. When the check arrives, you decide to split it, and each do the math separately. You calculate that $43 is the correct amount each of you should pay, and become quite confident in this claim. Your friend comes up with $45, and is quite confident in their answer. You have no independent reason to think your friend is worse at elementary addition than you are. How should you respond?

Intuitively, many philosophers have thought that you should become less confident that the correct answer is $43. After all, at least one of you has to be incorrect, and you have no reason to think that your friend is any less competent at basic arithmetic than you are. At least one of you has made a mistake, but for all you know the mistake is yours. Given this, it makes sense to be less confident in your own answer, where we understand this confidence in terms of degrees of belief (or credences).

More carefully, the intuition is that your degree of belief in the proposition "Half the check is $43" should decrease, and your confidence in "Half the check is $45 should increase." Alternatively, for those who prefer talking in terms of categorical belief, the disagreement should cause you to give up your belief in $43 being the correct answer, and should lead you to instead suspend judgment. Conciliationism generalizes from intuitions about cases like *Mental Math* to a general theory about the epistemically rational response to disagreement.

Conciliationist theories are those which (at least often) call for a subject to significantly change her belief state in response to the mere fact of disagreement. That is, the discovery of the disagreement itself is the evidence (or at least the impetus) that leads to the change in belief. More specifically, conciliationist theories call for a subject to change her beliefs to be significantly closer to those of her disagreeing interlocutor, at least in many cases of disagreement.

As I will use the term in this paper, a conciliationist view is one that will agree with the following general principle, or something close to it:

> **Conciliationism (CV):** If a subject has a certain credence c_1 with respect to p, and then learns an epistemic peer has a different credence c_2 towards p, then the subject should (often, if not always) adopt a new doxastic state c_3, which is *significantly* closer to c_2 (than c_1 is).[1]

Several aspects of the above principle need explanation. First, as I will understand things, the kind of disagreement mentioned in CV consists in learning that a peer has a different credence towards a relevant proposition.[2] Second, an epistemic peer is another person whose epistemic position is just as good as the subject's. This notion of epistemic position encompasses both the available evidence, and competence at evaluating evidence, relative to a subject matter. That is to say, a genuine peer has all the same evidence as the subject, and is just as good at evaluating evidence as the subject. Intuitively, in the mental math case, you and your friend are peers because you both have access to the check and a similar mastery of elementary arithmetic. Much ink has been spilled about the nature of epistemic peerhood, but this debate does not concern us here.[3] The skeptical results problem and self-undermining problem arise for conciliationism on any of the plausible ways of cashing out the notion of an epistemic peer. Moreover, the problems arise for any view that recommends conciliation in cases of non-peer disagreement, e.g., in cases where the disagreeing interlocutor is in an epistemic superior (or even inferior) position. Most conciliationist views will call for conciliation in these cases as well. For our purposes, then, "epistemic peer" can be read as any interlocutor with at least an approximately similar epistemic position.

---

[1] This version of the generalized CV principle is adapted from Elga (2010) and Matheson (2015a). I have framed CV in terms of credences, but a similar principle can be formulated for categorical or full beliefs. For ease of exposition, I will focus on the credence version of the principle, but both the two problems, and my proposed solution, will apply to both versions.

[2] Here I am following Lasonen-Aarnio (2013)

[3] For an overview of this debate, and some more nuance about the available positions, see Frances and Matheson (2018); N. L. King (2012).

Conciliationism is a category of views that contains a variety of more specific theories which specify how a subject should change her beliefs, and by how much. The most famous of these is the *equal weight view* (EWV) (Christensen 2007; Elga 2007). According to EWV, a subject is always required to significantly conciliate in circumstances of peer disagreement, by according their peer's judgment equal weight to their own (provided the subject does not gain evidence independent of the disagreement that the interlocutor is not really a peer).[4]

EWV has often been seen as an extreme view, as it calls for significant revision of credence in all cases of peer disagreement, and puts stringent independence requirements on dismissing an interlocutor's peerhood. Several philosophers have proposed more "moderate" views. These are meant to require less significant conciliation, and allow for "downgrading" purported peers on the basis of the very disagreement in question.[5] Moderately conciliatory views are committed to CV, as I have construed it. This is a helpful way of classification because the problems we will consider here, the skeptical results problem and the self-undermining problem, arise for both extreme and moderate versions of conciliationism.

Conciliationism's opposite is known as steadfasting, or the steadfast view (SV). Steadfast views claim that one should respond to disagreement by maintaining one's initial credence. That is, (in at least many cases) the mere fact of disagreement does not provide any reason to lower one's confidence or change one's belief state. In general, steadfast views will accept the following principle:

> **Steadfastism (SV):** If a subject has a certain credence c_1 with respect to p, then (at least often) the subject should maintain c_1 when she learns an epistemic peer has a different credence c_2 towards p.

There are also a number of different steadfast views, but their differences will not

[4]It turns out that a precise specification of exactly what this means that is compatible with Bayesianism is hard to formulate. See Jehle and Fitelson (2009); Lasonen-Aarnio (2013); Rosenkranz and Schulz (2015); **?**.

[5]For examples of moderate conciliatory views, see Kelly (2010) , Sosa (2010), Lackey (2008), and Worsnip (2014).

concern us.[6]

As stated, CV and SV are in opposition largely due to the fact that conciliatory views *at least often* call for significant conciliation, while steadfast views *at least often* call for no such conciliation. This characterization is admittedly a bit imprecise, in virtue of the term "often." However, this formulation counts moderately conciliatory views as as accepting CV rather than SV. This is helpful in the current circumstances, because the two problems we will discuss arise even for these moderate views.[7]

The skeptical results problem and the self-undermining problem are objections to conciliationism in general, in favor of steadfast views in general. The two objections are supposed to show that even those with conciliationist intuitions in cases like *Mental Math* should nonetheless give up conciliationist theories. My project here is to show that these objections fail, once we adopt the theory of endorsement. Some philosophers do not seem to share the conciliating intuitions from the mental math case, and some are motivated to be steadfasters for other reasons. The arguments here are not meant to persuade such philosophers to give up the steadfast view. Rather, they are simply meant to block these two, significant objections to conciliationism.

The kind of disagreement we see in the mental math case is often taken to work as a kind of *higher-order defeater* (Christensen 2010a; Lasonen-Aarnio 2010, 2014; Schoenfield 2016; Titelbaum 2015). The evidence provided by the disagreement in cases like mental math is evidence that the subject has made a mistake. This is evidence that something is wrong with the subject's beliefs, not evidence directly against the proposition in question (e.g., that half the check is $43). Disagreement is not the only kind of higher-order defeater. For instance, finding out that I might be suffering from hypoxia (i.e., oxygen deprivation) can give me reason to doubt my belief-forming

---

[6]For versions of steadfastism, see Kelly (2010), Titelbaum (2015). and Lasonen-Aarnio (2010, 2014). For additional background on the peer disagreement literature, see Christensen (2009), Matheson (2015b), and Frances and Matheson (2018).

[7]Note also that as stated, CV and SV concern cases where a subject discovers actual disagreement. There is a long-running dispute regarding whether merely potential disagreement should be treated the same, epistemically, as actual disagreement. For an overview of this dispute, see Matheson (2015b). I suspect that potential disagreement should not be treated as epistemically relevant in the same way that actual disagreement is. However, this dispute is yet another that won't concern us here. The two objections arise in either case, and the theory of endorsement will also provide a solution either way.

methods (Christensen 2010b). In this paper I will focus exclusively on disagreement. However, the solutions to the problems here will also help in resolving similar difficulties for dealing with other kinds of higher-order defeaters.

3.2.1 The Skeptical Results Problem for Inquiry

The first problem for conciliationism that I will address is the skeptical results problem.[8] I will be particularly concerned with this problem as it arises for cases of inquiry. To summarize, the problem is that according to conciliationism, we should be much less confident in our views than we seem to be when faced with systematic disagreement.

In contemporary, cutting-edge research fields across the sciences, humanities, and elsewhere, we find researchers who are committed advocates of their theories. Researchers defend their theories and remain committed to them over time. They base their future research on these commitments. They assert their theory's superiority, or even simply categorically assert its implications. In everyday discourse, such behavior is generally associated with belief and/or high credence in the the propositions in question.

At the same time, cutting-edge research fields are characterized by pervasive disagreement. Researchers are committed advocates of different, incompatible theories, and the practice of inquiry involves disputes between disagreeing researchers. Of course, there is often significant scientific consensus about various claims within a field. But these claims are not then the focus of cutting-edge research. Cutting-edge research occurs on questions where there is no consensus, and in these cases we find the relevant sort of disagreement.

The skeptical results problem for conciliationism arises from the tension between the committed advocacy of researchers and the mandates of conciliationism in light of the actual disagreement of researchers. According to CV, researchers should not believe or have high credence in their theories, given the prevalence of disagreement. But

[8] This problem has been pointed out by a number of philosophers, including Frances (2010, 2013), Goldberg (2013a), Matheson (2015b), and Sosa (2010).

researchers behave as though they do believe. So it looks like conciliationism has the implication that many (perhaps most) scientists, humanists, and other researchers are behaving irrationally (at least from an epistemic perspective). Conciliationism mandates skepticism about theories in cutting-edge research fields, and this skepticism means that researchers should not be committed advocates of particular theories.[9]

This seems counter-intuitive. Moreover, if this is right then conciliationism has the result that individual epistemic rationality is in significant tension with the collective goals of inquiry. The goals of collective inquiry are clearly supported by having individuals serve as committed advocates of various views, and by vigorous and productive debate and dispute. That such disagreement is beneficial to inquiry is both intuitive and supported by empirical evidence from psychology and the history of science.[10]

The skeptical results problem is that conciliationism mandates a skepticism about our theories which seems at odds with, and undercuts, our best research practices. Below, I will argue that we can solve this problem for conciliationism by appeal to the theory of endorsement. Before turning to a characterization of this theory, I will in the next section explore a second problem for conciliationism.

3.2.2 The Self-Undermining Problem

The self-undermining problem is perhaps the most pressing difficulty for conciliationism. Epistemologists have developed this objection in a number of ways, but since the solution I will offer for this problem is applicable to any of them, I will focus on two particular ways in which the problem arises.

---

[9]Some moderately conciliationist views are specifically motivated by the desire to avoid a version of the skeptical results problem, e.g., Sosa (2010). They appeal to private evidence (or other private epistemic bases) to protect important opinions about politics, religion, morality, etc. However, these moderate views are still subject the skeptical results problem for *inquiry*. Cutting-edge research is plausibly a paradigm case for when conciliation is required, even by the lights of moderate theories. This is because the relevant evidence is required to be public and shared, and the general level of reliability on answering the most advanced questions is known to be low for everyone, given the uncertain state of the field.

[10]For psychology see Mercier and Sperber (2011). For the history of science, see De Cruz and De Smedt (2013).

The basic idea of the self-undermining problem is that conciliationism's own recommendations, when applied to the epistemology of disagreement, recommend giving up (or significantly lowering credence in) conciliationism. That is, advocating conciliationism by taking part in the epistemology of disagreement gives you strong reason, by conciliationism's own lights, to give up conciliationism.

Here is a case we can use to illustrate the problem:

> **Connie the Conciliator:** Connie is an epistemologist. After reading the early literature on disagreement, she is attracted to conciliationism. She finds herself becoming a committed advocate of the view, exploring and defending it. Later, she is confronted with disagreement from some of her peers, who are steadfasters. As a conciliationist, she dutifully lowers her confidence in conciliationism, in light of this disagreement. She now has less than .5 credence in the theory, and certainly less credence than would be required to rationally believe the theory.[11]

There are two main versions of the self-undermining problem that have been proposed. The first version, which I will call *self-undermining justification*, is due originally to Plantinga (1999). The second version, which I will call *self-undermining inconsistency*, was proposed by Elga (2010) and Weatherson (2007). We can use Connie's case to illustrate each version.

The basic idea of the self-undermining justification problem is that adopting conciliationism causes a subject to lose her justification to believe conciliationism. The term "adopting" is meant to be a neutral term for being committed to a theory. This could either mean believing the theory, or as I will argue later, endorsing it. The self-undermining justification problem arises when we interpret adopting to mean believing. To see how the problem arises, let's consider the argument as it applies to Connie's case:

---

[11] Compare this example to cases from Elga (2010) and Weatherson (2007).

Self-undermining justification

1. Connie adopts CV, then meets several disagreeing peers.
2. CV (given the evidence of disagreement) recommends that she lower her confidence in CV to the point where it is too low for her to be justified in believing CV.
3. Connie should follow this recommendation and lower her confidence.
4. If she follows the recommendation, she is no longer justified in believing CV.
5. If Connie is not justified in believing CV, she should give it up (no longer be committed to it).
6. Thus, Connie should give up CV.[12]

Premise one is the assumption that Connie becomes committed to conciliationism, where usually this is interpreted as believing it. Premise two states that after Connie confronts the disagreement with her peers, conciliationism recommends lowering her confidence in CV. Because she is committed to conciliationism, according to premise three Connie should follow this recommendation. Because her confidence is too low in the theory, she is no longer justified in believing it (premise 4). Plausibly, if you are no longer justified in believing a theory, you should give it up, i.e., no longer be committed to it (premise 5). So Connie should not be committed to conciliationism.

If this argument is sound, this is a serious problem for conciliationism. It looks like adopting the theory is unstable: adopting it, in the current circumstances of disagreement, seems to lead one to be rationally required to give up the view. In other words, adopting the view is unstable rationally speaking. This looks bad for conciliationism.

The self-undermining inconsistency objection is perhaps even worse: it alleges that conciliationism is incoherent because it gives inconsistent advice in many circumstances in epistemology of disagreement. This can also be illustrated by appeal to the argument as it applies to Connie's case. Suppose that Connie proceeds as before, and has met several peers who disagree with her.

---

[12] See Plantinga (1999), Weatherson (2013), Decker (2014) and Matheson (2015a) for this version.

Self-undermining inconsistency

1. Connie adopts CV, then meets several disagreeing peers.
2. CV (given the evidence of disagreement) recommends that she lower her confidence in CV to the point that she has greater confidence in SV.
3. If Connie has greater confidence in SV than CV, then she should follow the rules prescribed by SV.
4. Thus, CV and Connie's evidence recommend following the rules prescribed by SV.
5. CV recommends both having a lowered credence in CV, and following SV (which recommends maintaining her credence).
6. CV gives inconsistent advice (because of 4).
7. One should not maintain a commitment to a theory which gives inconsistent advice.
8. Therefore, Connie should give up CV.[13]

Premise two here suggests that, in Connie's case, there is adequate disagreement to warrant Connie actually having less than 1/2 confidence in CV, and so having greater confidence in SV. This is a reasonably plausible situation for Connie to find herself in, given the current circumstances in the epistemology of disagreement.[14] Premise three is plausible if what one should do depends on the theory one has most confidence in (something I will take issue with shortly). Premise five follows from the different commitments in SV and CV. Steadfastism recommends something inconsistent with conciliationism in this case: SV recommends that Connie maintain her level of confidence in CV. CV, meanwhile recommends that Connie lower her confidence in CV. It's impossible for her to follow both recommendations. Since CV gives inconsistent advice that cannot be followed, Connie should give it up as a theory of what rationality requires her to do. One cannot be rationally required to do something impossible.

---

[13] See Elga (2010) and Weatherson (2013) for this version, which is inspired in part by Lewis (1971).

[14] Part of the point here is to grant the objector the strongest possible case for the objection. Later, I will show that Connie would be rational to maintain her commitment to CV, despite this.

If this argument is sound, then it at least appears that conciliationism is self-undermining, because adopting the theory results in receiving inconsistent advice.[15] This looks even worse for the view than the first version of the problem. Not only does adopting the view leave one unjustified, it turns out that adopting the view is incoherent.

Notice, however, that each of these versions of the problem relies on additional commitments, over and above a commitment to conciliationism. First, each requires that we treat "adopting" a theory as believing it. As I will argue later, this assumption should be abandoned. Second, and more importantly, each of these arguments appeals to a commitment to a certain kind of *enkratic principle*: a principle that constrains what a subject should do (or believe), given what they believe they should do (or believe). Enkratic principles rule out the rationality of *akrasia*: believing that one should do something, and yet not doing it. The kinds of enkratic principles at issue here involve prohibiting akratic beliefs: believing something (to a certain degree), while (justifiedly) believing that one should not so believe.

The self-undermining justification objection relies on what we can call the *Unjustified Theory Principle*:

Unjustified Theory If a subject S is not justified in believing (or would not be justified in believing) a theory, then S is rationally required to give up (not be committed to) that theory.

This principle is a general expression of the principle in premise 5 of the self-undermining justification problem. Without this principle, the conclusion does not follow. It is an enkratic principle, in that it offers advice about how to act in accordance with what one is justified in believing. I will argue that although this principle seems intuitive, it is in fact false. Being committed to a theory is rationally compatible with being unjustified in believing the theory. This is because the proper way to be committed to a

[15] Note that the claim here is not that the view is self-refuting. Adopting conciliationism is not incompatible with the truth of conciliationism. The problem, in both versions, is an epistemic problem: one cannot rationally adopt the theory, because doing so undermines the theory's justification, and results in the subject getting inconsistent advice. This is not the same kind of self-refutation that seems to arise in common paradoxes, like the Liar Paradox. See Matheson (2015a) and Decker (2014).

theory during inquiry is to endorse it.

The self-undermining inconsistency problem is committed to a different kind of enkratic principle. As stated, the argument relies on what we can call the *Most Credence Principle*:

Most Credence If a subject has the most credence in a theory, then the subject should follow any rules prescribed by that theory.

This is an enkratic principle that connects the subject's credence in competing theories about rules, with whether they should act on those rules. In particular, it requires subjects to follow rules which appear in theories they have higher confidence in than in any competitor theory.

The Most Credence Principle underwrites the third premise in the inconsistency argument. Premise three does not follow without a commitment to an enkratic principle of this kind. This is because conciliationism is a theory that only prescribes how subjects should change their beliefs in light of disagreement. Conciliationism itself does not include any advice about when to follow rules prescribed by theories one has high credence in.[16] CV and Connie's evidence imply that she should lower her confidence in CV and raise it in SV. CV itself says nothing about how to act upon rules contained in theories like SV. It is only in combination with the Most Credence Principle that Connie's newfound high credence suggests that she should now follow the dictates of steadfastism. The self-undermining inconsistency argument is only valid with this suppressed premise. Without the Most Credence principle, there is no inconsistency.[17]

---

[16]Lasonen-Aarnio (2014) also briefly notes that reasons to doubt a theory are not always thereby reasons to stop following the theory, and suggests that this is a problem with formulations of the inconsistency objection. However, she doesn't go on to analyze which extra assumptions are necessary to make the argument work.

[17]Weatherson (2013) and Elga (2010) (in a footnote) both suggest an alternative formulation of the self-undermining problem. This alternative version suggests that CV is self-undermining if it suggests *any* lowering of credence in itself. The argument actually requires commitment to a different enkratic principle, which we can call the *weighted average principle*. This principle is that one should follow rules which are derived from a weighted average of rules contained in all the theories one has some credence in (with the weights provided by the credences). It is not clear to me that this principle is generally applicable: it seems like it only makes sense where the rules to be followed can be expressed using real

Each version of the self-undermining argument relies on a particular enkratic principle. I will argue that each of these principles is false. One can adopt a theory even if one would be unjustified in believing it. One should not always follow the rule that one has most credence in. In order to motivate these claims, I will present a general theory about inquiry which denies them. According to this theory, the appropriate kind of commitment to a theory is an attitude called endorsement. In the next section, I will present this theory of endorsement, before returning to argue that the theory provides strong motivation for denying the two enkratic principles.

3.3 Endorsement

So far, I have described conciliationism and two of its most pressing problems. In order to solve these problems, I am now going to present the theory of endorsement.

As I have suggested above, we often find researchers who remain committed advocates for their theories over time. Such inquirers advocate for their theories, assert their superiority, and defend them doggedly against difficult objections. Moreover, that researchers act this way is a good thing: such practices are constitutive of healthy inquiry. But again, in ordinary contexts, such commitment and advocacy is explained by the fact that people believe the claims in question. However, according to conciliationism, belief and even high credence in one's own theory would be irrational in the face of the systematic disagreement researchers confront in cutting-edge inquiry. In fact, there are additional reasons to doubt that belief is warranted. In cutting-edge fields characterized by disagreement, the available evidence underdetermines which theory to choose. Moreover, most theories which have been proposed have not ultimately been accepted, but have been discarded as false (this is a version of the pessimistic meta-induction (Psillos 1999)).

Despite the usual connection between belief and committed advocacy, and despite the apparent irrationality of belief in the relevant contexts, we do not want our theory

---

numbers (and thus can be numerically averaged), and this won't always be the case. Regardless, the same endorsement solution applies to this version of the self-undermining problem as the one appealing to the *Most Credence Principle*. I will focus on the version of self-undermining inconsistency which appeals to *Most Credence*, as I think it is more plausible, and doing so will ease the exposition.

of epistemic rationality to count researchers as irrational for being committed advocates of their views. Nor do we want it to prohibit committed advocacy, which is beneficial to inquiry. There is thus a tension between our theory of individual epistemic rationality, and the goals of collective inquiry.

In order to resolve this apparent tension, I propose that we recognize a distinct doxastic attitude, one which is appropriately governed by rational norms concerning collective inquiry. I call this attitude *endorsement*, and the norms which govern it *inclusive epistemic rationality*.

As I will argue below, this theory of endorsement will provide solutions to both problems for conciliationism discussed in the previous section. In the rest of this section, I will first provide a characterization of endorsement, followed by a brief explanation of inclusive epistemic rationality.

3.3.1 The Nature of Endorsement

Endorsement is a propositional, doxastic attitude. It embodies the resilient commitment and advocacy that researchers should have toward their theories in a field of healthy inquiry. The following characterization is a part of the attitude's functional profile that we can use to distinguish it:

> **Endorsement:** Endorsement is a doxastic propositional attitude. It is an attitude of resilient commitment and advocacy. S endorses p in a research domain d only if:
>
> 1. S is disposed to assert that p, or otherwise express commitment to p (in d).
> 2. S takes herself to be obligated to defend p (in d).
> 3. S treats p as a premise in her further reasoning (in d).
> 4. S shapes her research program in d (in part) based on p.
> 5. S is resiliently committed to p (in d).
> 6. In endorsing p, S aims to promote healthy inquiry.

This characterization is not meant to be a traditional definition of a concept (i.e.,

a set of necessary and sufficient conditions). Instead, it simply provides some characteristic markers as a way of distinguishing the attitude from other propositional attitudes (e.g., desire, hope, or fear) and other doxastic attitudes (e.g., belief, credence, or scientific acceptance).[18]

There are a few things that need to be clarified in this characterization. First, a subject endorses a proposition within a particular research domain. This makes endorsement a "fragmented" attitude, meaning it is compartmentalized to one part of the subject's mental life, rather than being a global feature of their mental state.[19] This fragmentation helps explain the way subjects can behave as committed advocates within a research domain, but not be willing to accept the theory (or gamble on it for high stakes) outside of the domain.

Endorsement is related to other kinds of doxastic attitudes, such as categorical belief and credence. It is an attitude toward the truth of a statement or proposition. However, it differs from each of these other attitudes in a variety of ways. The last three of the necessary conditions in particular allow for the attitude to be distinguished from categorical belief. These three conditions require dispositions to act in ways that are not required for a belief. One need not be obligated to engage in a research program on the basis of any belief one has. Endorsing a theory in the context of inquiry, however, involves a practical commitment to engage in such research.[20]

The resilient commitment requirement in condition (5) also serves to distinguish endorsement from categorical belief. The sense of resiliency here involves a maintenance of the commitment in the face of contrary evidence. In this sense, belief is less resilient than endorsement. When faced with strong contrary evidence, such as a

---

[18]This notion of endorsement is inspired by acceptance/belief distinction, especially by the work of L.J. Cohen (1989a), Levi (1980), Maher (1993), and Van Fraassen (1980). But there are many differing notions of acceptance in the philosophical literature. To pick out the particular, provisional and epistemic notion I am interested in, I have chosen the name "endorsement." Recently, several philosophers have recognized the need for a provisional acceptance attitude of some kind, e.g., Goldberg (2013a), McKaughan (2007), and **?**. Elgin (2010) appeals to Cohen's notion of acceptance as a way of dealing with the skeptical results problem. I think endorsement does the best job of playing this provisional acceptance role.

[19]For more on fragmentation, see Egan (2008a) , Elga and Rayo (2015) and Rayo (2013).

[20]This point is inspired by discussion about the difference between belief and acceptance in Whitt (1990).

purported counter-example, a rational subject should give up the belief. However, a researcher's endorsement of a theory often survives the discovery of significant contrary evidence. As I will suggest in the next sections, this resiliency in the face of contrary evidence is rational, and beneficial for collective inquiry.[21]

Endorsing P is compatible with suspending judgment (also called "withholding belief") about whether P. Generally, endorsing a theory will involve suspension of judgment on the question that the theory provides an answer to.[22] Importantly, endorsing P is compatible with having a particular credence in P. On my account of rational endorsement, what it is rational to endorse will in part depend on one's credences.

Endorsement is the appropriate attitude for researchers to take toward the theories they pursue as committed advocates. It is a provisional attitude, taken toward the theory that seems best to pursue to the researcher in question. Categorical belief, on the other hand, is an attitude appropriate toward the answer to a question one takes to be settled. Because of this difference in roles, belief and endorsement are governed by different epistemic standards. For instance, one cannot rationally (or justifiedly) believe a proposition P when one takes $\neg P$ to more likely: in other words, you should not believe P if your credence in P is less than .5. Moreover, you should not believe P when you take a competitor theory Q to be more likely to be true, i.e., when your credences have it that $Pr(P) < Pr(Q)$. On my account, endorsement can be rational in both of these circumstances. This is due to its nature as a provisional attitude taken during inquiry.

With an account of the attitude on the table, we can turn now to a characterization of the norms that govern it.

---

[21] This part of the functional profile for endorsement builds in some normative requirements. I think this will be inescapable in giving an account of a doxastic attitude. Part of the role that such an attitude has in a subject's mental economy is its relationship with other attitudes, and these are governed by rational norms. If a subject's attitude toward P violates too many of the norms of belief, and does so in a systematic way, then that attitude will simply fail to play the belief role (or be in the "belief box") at all. However, if one prefers to avoid normative talk in delineating a functional role, they could simply treat the norms as generalizations, and the effect for my purposes will be the same.

[22] For this notion of suspension, see Friedman (2017).

3.3.2 Inclusive Epistemic Rationality

Part of what distinguishes endorsement from belief and credence is the distinctive kind of epistemic reasons to which endorsement is sensitive. I call these *extrinsic epistemic reasons*, and the normative framework that includes such reasons *inclusive epistemic rationality*. Sensitivity to extrinsic epistemic reasons is an important aspect of the theory of endorsement's solution to the skeptical results and self-undermining problems.

Inclusive epistemic rationality is inclusive of both intrinsic and extrinsic epistemic reasons.[23] Intrinsic epistemic reasons are reasons which are about, or indicate, the truth of the proposition in question. They are reasons to think that a proposition is true. If Q is evidence for P, then Q is an intrinsic epistemic reason (for P). Other intrinsic reasons are things which serve as necessary conditions for truth: e.g., that a theory is consistent is (some) reason to think it is true, and therefore is an intrinsic reason. Intrinsic epistemic reasons indicate, or point to, the truth of the proposition in question.[24]

Extrinsic epistemic reasons are reasons concerning what will promote healthy inquiry. They are reasons about what will lead to getting more truths, or more knowledge, in the long run. Such reasons might concern only the productivity of an individual engaged in research alone. However, extrinsic epistemic reasons can also be reasons concerning the promotion of the goals of collective inquiry.

Extrinsic epistemic reasons go beyond reasons to think a theory is true. They are reasons in favor of a proposition that concern how taking some attitude (or action) regarding the proposition will affect inquiry. For instance, the testability of a theory is often taken to be an important feature for the purposes of scientific inquiry. This

---

[23] This particular intrinsic/extrinsic distinction is originally due to Steel (2010), though I have developed it differently than he does in that paper. Lougheed and Simpson (2017) are concerned with the same kind of distinction.

[24] I am using the normative language of reasons, but his is largely for convenience. The framework presented here is compatible with using "ought" language, or "value" language. I also intend to be neutral between different underlying theories of justification and rationality. Whatever background theory one has about these concepts (or relations), it ought to be compatible with our best account of the epistemic norms for individuals engaged in collective inquiry.

is not because testability indicates truth (most testable theories have turned out false) (Steel 2010). What testable theories have going for them is their suitability to inquiry. Because of this, that a theory is testable is an extrinsic epistemic reason to pursue it. Other extrinsic epistemic reasons concern the division of labor in collective inquiry: if a theory (or paradigm, or project) has very few people working on it, this is an extrinsic epistemic reason to endorse and pursue it. This is because collective inquiry is likely to go better if there is a better distribution of cognitive labor.[25]

On my account, extrinsic epistemic reasons are not merely pragmatic reasons. They deserve to be considered "epistemic" because they are about inquiry, about the pursuit of knowledge. But they are not the kinds of reasons that have been typically treated as epistemic, as they do not indicate the truth of the proposition they relate to. Nonetheless, I think it is worth distinguishing between reasons concerning the promotion of practical ends (like becoming rich and famous) and reasons which concern successful inquiry. Extrinsic epistemic reasons are not about merely practical benefits, but about long-term or collective epistemic benefits. For instance, if a researcher chooses to endorse a theory because it has too few defenders, this is a reason that has to do with promoting healthy inquiry, not with fame or fortune. It is useful, theoretically, to treat these reasons as distinct from practical reasons.

Of course, researchers may (and most likely will) also be motivated by pragmatic considerations. However, it seems a common occurrence that researchers are also motivated by a desire to contribute to healthy inquiry. Researchers who are thus motivated should not be evaluated as irrational or unjustified by our best theory of epistemic rationality. Rather, our theory should vindicate and encourage researchers who are sensitive to reasons concerning the overall health of collective inquiry. Thus, we need a category of genuinely epistemic reasons which go beyond reasons to think a theory is true. The category of extrinsic epistemic reason plays this role.

One distinctive feature of endorsement is that it is sensitive to both intrinsic and extrinsic epistemic reasons. This is distinctive because it is highly plausible that belief

[25]See Kitcher (1990) and Strevens (2003).

(both categorical and partial) is not sensitive to extrinsic epistemic reasons, but only to intrinsic ones. This is illustrated by the common intuition that it is epistemically irrational to believe in a way that does not fit one's evidence, or in a way that is known to be unreliable. Beliefs should fit the evidence, or be reliably formed (preferably both).

The rational insensitivity of beliefs to extrinsic epistemic reasons is also illustrated by the common intuitions to epistemic bribery cases.[26] It looks generally inappropriate to take on a belief that is likely to be false, or is not supported by one's evidence, in order to gain other true beliefs. If one is convinced of atheism on the evidence, it is epistemically irrational to believe in God, even if that belief will provide one with a great deal of research funding and thus many more true beliefs in the long run. This is an extrinsic reason to be a theist, and thus does not support belief in theism.

Endorsement, on the other hand, is sensitive to reasons having to do with healthy inquiry, and is thus an appropriate attitude to take toward theories which are promising, or have too few defenders, but which the subject does not take to be highly likely to be true. What it is rational for a researcher to endorse depends on the balance of intrinsic and extrinsic epistemic reasons: that is, on inclusive epistemic rationality. Sometimes, the extrinsic epistemic reasons to endorse a theory will outweigh the intrinsic epistemic reasons in favor of its competitors, and it will be epistemically rational for a researcher to endorse a theory which she takes to be less probable than its competitors.

The sensitivity of endorsement to extrinsic epistemic reasons also helps to explain the resilience of endorsement in the face of contrary evidence. A researcher who endorses a theory that is faced with a difficult objection, or purported counterexample, might rationally lower her credence significantly in the theory. Still, her extrinsic epistemic reasons may outweigh this loss of credence, and make it rational for her to maintain her endorsement. For instance, she might be one of very few researchers working on the theory, or it might be a significant cost to change research programs

[26] For these cases see Firth (1981), Jenkins (2007), Berker (2013), and Greaves (2013).

(based on available laboratory space or research equipment). As long as the evidence does not completely rule out the theory she endorses, the extrinsic epistemic reasons to continue pursuing her theory can outweigh the intrinsic reasons against the theory.

Rational sensitivity to extrinsic epistemic reasons helps the practice of endorsement promote healthy inquiry. Researchers will have reasons to endorse different theories from one another, even when in agreement about the available evidence. This will promote a useful distribution of cognitive labor, and will also lead to beneficial kinds of disagreement that helps researchers reason more effectively, and which promotes better evidence collection.[27] Sensitivity to extrinsic reasons will also help promote resilient commitment, as extrinsic reasons to maintain a commitment will remain even in the face of difficult contrary evidence. Thus, endorsement and its inclusive epistemic rationality provide significant benefits to inquiry.

These features of endorsement will allow the theory of endorsement to solve the two problems for conciliationism raised above.

3.3.3 The Value of Endorsement

Endorsement is the doxastic attitude which embodies the committed advocacy that researchers rationally take to their own theories during the course of inquiry. It is governed by inclusive epistemic rationality, and sensitive to extrinsic epistemic reasons. Due to this sensitivity, a practice of rational endorsement will promote a number of valuable features for a field of inquiry. These benefits to inquiry help to justify the current practice of at least some researchers, and the adoption of such a practice of endorsing theories where it does not already obtain. This provides the endorsement solution to conciliationism's problems with significant independent motivation. In order to support this claim, I want to briefly mention some of these beneficial features.

The first valuable feature of an inquiry that endorsement promotes is a motivated commitment of researchers. The basic idea is that researchers who are committed to, and identified with, a particular theory or paradigm will be motivated to explore,

[27] For some of the empirical evidence supporting the benefit of disagreement claims, see Mercier and Sperber (2011), and De Cruz and De Smedt (2013).

support, and defend a theory to the best of their ability. A researcher who endorses a theory within a domain of inquiry has a horse in the race. Such a researcher will behave in ways that benefit inquiry. They will commit significant time and energy to exploring and defending the theory. They will be motivated to develop and promote the theory, and to give it its best possible defense. A researcher who endorses a theory will be motivated to explore all implications of the theory, in order to find new ways of improving and supporting it in the face of competition. Moreover, they will continue defending the view even in the face of difficult objections and contrary evidence, and will avoid premature abandonment of theories which remain live possibilities.

The practice of such motivated research will contribute to the health of inquiry in a variety of ways. It will lead to a better distribution of labor among researchers, will help avoid premature scientific consensus, and will help to avoid costly switching between research programs when evidence for various theories is accumulated unevenly (that is, in a way that does not reflect the totality of evidence that will eventually be accumulated).

A second, related valuable feature of inquiry promoted by endorsement is a kind of beneficial disagreement. This is because it is often rational to endorse a theory which does not have the balance of evidence in its favor, if the extrinsic epistemic reasons for a researcher to endorse it are strong enough. Since extrinsic epistemic reasons will be different for each researcher, this will lead to researchers rationally endorsing different positions. Since endorsement is an attitude of committed advocacy, this will also lead these researchers to advocate and argue for what they endorse, and thus lead to disagreement in the field.

There is significant reason to think this kind of reasoned dispute and debate within a field will be a benefit to collective inquiry. First, the value of this is intuitive, and such debate has been regarded as valuable historically. This historical positive evaluation is enshrined in a variety of institutions, including democratic governing bodies, academic fields, and adversarial judicial systems. Intuitively, then, it has been thought valuable to encourage a "marketplace" of ideas, where the best idea is expected to emerge from rigorous debate among and between advocates for each view.

In addition to the intuitive evidence, and the fact that it is traditionally attested there is empirical evidence of the value of disagreement and debate from both psychology and case studies in the history of science. In psychology, a number of studies support the idea that groups characterized by disagreement outperform both individuals and more initially-harmonious groups when engaged in reasoning and problem solving tasks (Mercier & Sperber 2011). Several philosophers have also pointed to a particular case in paleo-anthropology as supporting the merits of disagreement within a field of inquiry (e.g., De Cruz and De Smedt (2013)). In the debate about the purported discovery of a new hominid species, *H. floresiensis*, researchers have been led to produce a greater abundance of evidence, draw in evidence from other fields, and to more carefully interpret and evaluate the evidence produced by other researchers, because of the disagreement between two factions. Both the psychologists and those studying the *H. floresiensis* case have pointed out that such disagreement also benefits group inquiry by harnessing confirmation bias for good ends: it leads to a distribution of labor in production of new evidence, as each side of the debate seeks new evidence in support of the theory they endorse.

Thus, there is significant intuitive, historical, and scientific evidence to think that disagreement is beneficial for inquiry. The practice of endorsement encourages this beneficial disagreement.

These two benefits, motivated researchers and beneficial disagreement, make the practice of endorsement valuable for collective inquiry. Such a practice is thus worth adopting. The behavior of many researchers who remain committed advocates of their own theories, even when believing those theories would be irrational, is also vindicated by understanding them as engaging in this valuable practice of endorsing their theories.

3.4 Endorsement and the Skeptical Results Problem for Inquiry

So far, I have presented conciliationism, and introduced the two main problems for it that are at issue. In the last section, I introduced the resources necessary for solving

these problems: the attitude of endorsement, and the inclusive epistemic rationality that governs it. With these things on the table, we can see how endorsement solves the skeptical results problem and the self-undermining problem. In this section, I will present the solution to the former, and in the next section I will turn to consideration of the self-undermining problem.

Recall that the skeptical results problem, at its root, is that researchers are confronted with pervasive disagreement during inquiry, and conciliationism calls for significant lowering of confidence in such cases, with a concomitant suspension of judgment in the theories under dispute. Given this required low credence and suspension, and given that committed advocacy of theories should be understood in terms of believing those theories, its seems such committed advocacy is unwarranted. Conciliationism thus seems to call for skepticism and lack of commitment to a theory. Disagreement during inquiry, when you are a conciliationist, forces you to give up your commitment to a particular theory or research program (assuming such a commitment is only appropriate towards things you believe).

As I have argued above, however, the committed advocacy displayed by researchers toward their theories should not be understood as involving belief at all. Instead, endorsement is the attitude of committed advocacy that researchers should take toward their theories. Crucially, endorsement is compatible with low confidence and suspension of judgment. It can be rational to endorse a theory even when one has a low credence in the theory, provided there are good enough extrinsic epistemic reasons in favor of endorsing it. A theory which you have low credence in can be the best one to endorse, if it has extrinsic epistemic reasons favoring it. Thus, it can be rational to endorse a theory where it would be irrational to believe it.

Endorsement serves as the rational attitude which underwrites a researcher's exploration, defense, and advocacy of a theory. This can be the case even after the researcher discovers disagreement with her peers and duly lowers her credence in a theory, just as conciliationism requires. In a cutting-edge research field which has pervasive disagreement, conciliationism will require a researcher to lower her credence in her own favored theory. This will furthermore require her to give up her categorical

belief in the theory. But this lowered credence is straightforwardly compatible with the endorsement of the theory. Endorsement is an attitude of committed advocacy and pursuit. If an endorsement is rational, this justifies the researcher in continuing to pursue, explore, advocate, and defend her theory.

The theory of endorsement thus explains and vindicates researchers who remain committed advocates of their theory, at least when engaged in research. A practice of endorsement within a research field will promote a number of valuable features of inquiry (e.g., resiliency, beneficial disagreement, and division of cognitive labor), without conflicting with conciliationism.

The theory of endorsement solves the skeptical results problem by showing how researchers can be justified in continuing to be committed advocates of their theories even in the face of disagreement. It does not, however, suggest that researchers should believe their theories, or act on them outside of the practice of inquiry itself. This might seem like a limitation of the solution. However, I think that this is a feature and not a bug. In many cases, disagreement with epistemic peers should cause one to lower confidence, and to suspend judgment. This intuitive fact is what motivates conciliationism. A solution to the skeptical results problem that gave this up entirely would fail to preserve conciliationism. Instead, what we want is a solution which respects this intuition, while at the same time vindicating researchers. The theory of endorsement provides this kind of solution, and so its limitation to the domain of inquiry is no objection.

Notice that one need not be a conciliationist to be moved by the arguments in favor of endorsement. As I suggested above, there are a variety of reasons beyond disagreement to think that endorsement is the appropriate attitude to have toward a theory in cutting-edge inquiry. These reasons include a version of the pessimistic meta-induction, the underdetermination of theory by evidence, and the benefits a practice of endorsement provides for inquiry. It is true that the theory of endorsement is also partially motivated by cases of disagreement. But this is just a shared motivation with conciliationism. A commitment to CV itself is not what is motivating endorsement. Thus, the theory of endorsement is not merely something designed to

solve this very problem; it has independent motivation. This is a significant boon to endorsement's usefulness in solving the skeptical results problem for conciliationism.

The endorsement solution to the skeptical results problem works for two reasons. First, endorsement is a distinct attitude, one which the rational norms prescribed by conciliationism do not directly constrain. Second, endorsement is sensitive to extrinsic epistemic reasons which justify maintaining the attitude even in the face of lowered credence due to contrary evidence (in this case, evidence provided by disagreement). In the next section, I will argue that these same features also allow the theory of endorsement to solve the self-undermining problem.

3.5 Endorsement Defuses Self-undermining

As I argued above, the two versions of the self-undermining problem for conciliationism require appeal to additional assumptions beyond commitment to CV. The endorsement solution to self-undermining works by providing strong, independent motivation for denying these additional commitments. I have argued that we should understand "adopting" a theory to mean endorsing it, and I will now argue that the rule-following enkratic principles introduced above are false.[28] I will consider each of the two versions of the problem in turn.

According to the self-undermining justification argument, adopting conciliationism leads to a loss of justification for believing conciliationism, which is incompatible with continuing to be committed to the theory. This argument only works because its fifth premise is supported by *Unjustified Theory Principle*:

Unjustified Theory If a subject *S* is not justified in believing (or would not be justified in believing) a theory, then *S* is rationally required to give up (not be committed to) that theory.

[28] I will not argue that all enkratic principles are false, though my view is compatible with this position. However, one plausible way to interpret CV is that it is, itself, an enkratic principle (because it gives advice about how to change one's beliefs in light of learning they might well be mistaken). I will simply argue against rule-following principles like *Unjustified Theory* and *Most Credence*, which are required to make the self-undermining argument work.

According to the theory of endorsement, this principle is false. The attitude of commitment towards conciliationism that it is appropriate for a subject to have is endorsement, not belief. It can be rational to endorse a theory even when one would be unjustified in believing it. Rationally endorsing a theory requires less evidential support than justified belief. The extrinsic epistemic reasons that bear on whether to endorse a theory can provide strong support for endorsing it, while providing no support for believing it. Indeed, one of the main benefits of endorsement is that it vindicates commitments to theories which have inadequate evidential support to justify belief in them. Committed advocacy of a theory is beneficial to inquiry, so we have good reason to positively evaluate researchers who engage in this practice, and to actively encourage such a practice.

Once we recognize the value of endorsement, it is clear that we should not be committed to the *Unjustified Theory Principle*. Without this commitment, the self-undermining justification argument is invalid. The reason why the argument seemed compelling to many people is that *Unjustified Theory* is intuitive and plausible, as long as we think of belief as the only kind of commitment one could take toward a theory. However, once we recognize endorsement, we can see the principle is false.

Recall *Connie the Conciliator* from section 3.2.2. Connie's case can help illustrate how this solution works. Connie endorses conciliationism. Now, after engaging with the literature and discovering the actual disagreements with some peers in the epistemology community, CV requires her to lower her confidence in CV. Now she has a confidence too low to be compatible with justified belief. But that's fine! She doesn't believe her theory in the first place; she endorses it. And endorsement is compatible with a relatively low confidence. Connie has good extrinsic epistemic reasons to maintain her endorsement. Doing so creates a better distribution of labor in epistemology, promotes beneficial disagreement, and helps avoid premature consensus. So Connie is well-justified in endorsing the theory, even though she would be unjustified in believing it. Her lack of justified belief is no longer reason to give up the theory. Thus, Connie's justification for being a conciliationist is not undermined. The self-undermining justification problem is defused by endorsement.

The self-undermining inconsistency objection admits of a similar solution. According to the inconsistency version of the problem, conciliationism offers inconsistent advice in cases of disagreement, since it recommends increased confidence in a competitor that offers conflicting advice. Since CV recommends greater confidence in steadfastism, it thereby also recommends steadfasting as a rule. And it is impossible to both stay fast and conciliate.

As I argued above, however, the inconsistency problem also requires a commitment to an additional enkratic principle, over and above commitment to CV. It requires commitment to the *Most Credence Principle*:

Most Credence If a subject has the most credence in a theory, then the subject should follow any rules prescribed by that theory.

This is a special kind of enkratic principle that connects the degree of confidence one has in a theory, with whether one should act in accordance with rules or prescriptions contained in that theory.

The *Most Credence Principle* is an initially attractive principle for governing our decision-making under uncertainty. However, the theory of endorsement gives us strong, independent reason to deny it. The principle is incompatible with the kind of resilient commitment that endorsement involves, which provides significant benefits to inquiry. It would lead to subjects giving up theories too quickly, and without adequate exploration and defense. This is because, in cutting-edge fields of inquiry, new evidence is constantly being accumulated. New evidence can cause fluctuations in which theory a researcher has most credence in. This, in turn, could lead to fluctuations in the rules subjects follow, which will lead to incoherent, unstable fluctuations in what theory a subject is committed to. Thus, it will lead to a failure of resilience. In short, subject's who obey the *Most Credence Principle* will fail to be resilient in their endorsements, and this will be harmful for inquiry.

Without appeal to some enkratic principle like the *Most Credence Principle*, the self-undermining inconsistency argument does not follow. All the versions of such principles which have been appealed to in setting up self-undermining objections are incompatible with healthy inquiry (because they are incompatible with endorsement).[29] Thus, the argument for inconsistency is unsound.

The endorsement solution to self-undermining works by providing independent reason to deny the enkratic principles used in the objection. One might worry that there is nonetheless some undiscovered enkratic principle that would support the self-undermining argument. However, I think there is good reason to doubt that this is the case. It's hard to see what principle would both support self-undermining and allow for adequate resilience. Moreover, for those who are inclined to think there must be *some* rule-following enkratic principle which governs inquiry, the theory of endorsement itself motivates such a principle.

The endorsement framework prescribes a significant commitment to base one's research program on the theory endorsed. This involves giving the theory a full exploration, which includes investigating the implications of following the rules contained in the theory. This motivates a different kind of enkratic principle for rule following:

Endorsement Principle In a research domain d, if a subject endorses a theory, then (while engaged in research in d) the subject should follow any rules prescribed by that theory.

This principle is implied by endorsement's role in shaping a subject's research program. It is part of what it means to pursue and explore a theory, by investigating all of its consequences. Some of those consequences will concern how research is conducted. Thus, the Endorsement Principle helps make endorsement beneficial to inquiry. This provides significant independent motivation for this principle, as it is derived from

[29] As noted above in footnote 17, Weatherson (2013) and Elga (2010) (in a footnote) both suggest an alternative formulation of the self-undermining problem. This version appeals to a *Weighted Average Principle*. This principle would lead to even more fluctuation, and less resilience, than the *Most Credence Principle*. It is thus also incompatible with endorsement, and would lead to similar significant harms for inquiry. Therefore, the same solution applies to that version of the problem.

aspects of the endorsement theory which were not designed merely to solve the self-undermining problem.

The solution to self-undermining, augmented by appeal to the endorsement principle, can be illustrated by returning to Connie's case.

Connie endorses conciliationism. When she is faced with disagreement about disagreement, CV recommends lowering her confidence in CV, and raising her confidence in SV. According to the endorsement principle, she should follow this recommendation and lower her credence in CV (and raise it in SV). But having higher credence in SV than CV is compatible with her continued endorsement of CV. After all, she has a variety of extrinsic epistemic reasons in favor of maintaining her endorsement of conciliationism. Since she continues to endorse CV and not SV, the endorsement principle does not recommend that she follow rules prescribed by SV. Therefore, conciliationism does not give conflicting advice in her case. According to CV and the theory of endorsement, Connie should continue endorsing CV, and following its prescriptions to conciliate in the face of disagreement. Thus, Connie can consistently endorse CV.

The picture of disagreement that emerges from this discussion looks like this: Connie and her peers disagree in *credence*, and it is this disagreement that CV prescribes a response to. Notice that it is not the disagreement in *endorsement* that requires Connie to lower her confidence. This means that Connie must be able to get information about what her peers' credences are. However, this is already required by any theory which embraces CV. Recognizing that a practice of endorsement is in place may actually help researchers like Connie to disambiguate what information is being provided by bare assertions of theories within research domains, and motivate them to inquire more carefully about her peers' level of confidence. While it is true that researchers will need to be careful to acquire information about their peers' credences, and not just what they endorse, this is not a novel requirement introduced by the theory of endorsement.

Before discussing two significant objections to the endorsement solution to the self-undermining problem, I want to briefly mention one of its strengths. There is no space here to compare this solution in detail to other attempted solutions. However, the

advantages mentioned here go some way toward making the case that this is the best solution on offer. The advantage of the solution is the fact that it is embedded in an independently motivated research program.

The theory of endorsement offers a number of benefits beyond its solution to problems for conciliationism. We need it to properly explain and to vindicate the behavior of many highly effective researchers. There are a number of additional reasons, beyond disagreement, to think that belief cannot play the role of the attitude of pursuit during inquiry. Moreover, as was argued above, the practice of endorsement confers a number of benefits to a field of inquiry. In addition, the theory helps solve a number of separate problems in social epistemology and the general philosophy of science. It promotes an appropriate distribution of cognitive labor, it helps avoid premature consensus, and it contributes to our understanding of theory change and theory pursuit. Inclusive epistemic rationality is also poised to pay dividends in social epistemology on its own, by giving us an account of epistemic reasons for action.

Thus, the endorsement solution to self-undermining is embedded in an independently motivated research program. It also solves the problem in a way that respects the intuitions that motivate conciliationism, without carving out a special, ad hoc exception for the theory of conciliationism itself. These are significant advantages of the solution.

3.5.1 Objections

I have argued that the theory of endorsement provides a solution to both the skeptical results problem and the self-undermining problem for conciliationism. It seems to me that the skeptical results problem for inquiry is pretty definitively laid to rest by this solution. One might have other objections to the theory of endorsement (and I'm certain some do), but if something like that theory is accurate, then the skeptical results problem is solved. However, some might have lingering doubts about the solution provided to the self-undermining problem. In this section, I will try to allay these worries.

There are two potential objections I want to address here. The first is that the

self-undermining worry is something different than my analysis above suggests. The second is that, even if my analysis is correct, there is a part of the initial objection remaining. I will treat each of these concerns in turn.

The first worry is that my reconstruction of the self-undermining argument above fails to capture the real thrust of the objection. Someone pressing this worry might suggest that the true self-undermining problem arises earlier in the argument: that conciliationism does (or even merely can) call for lowering one's credence in conciliationism. It is the lowered credence, in the first place, that is the problem. The thought is, it is a bad feature for any theory to have that following its dictates rationally requires lowering confidence in that very theory. Thus, it is claimed, conciliationism has this problem even before one considers whether there is some categorical attitude one can consistently hold towards it. Merely the fact that one can be rationally required to lower confidence in a theory, by its own lights, is the essence of the problem.

I do not think this is really the essence of the self-undermining problem. At the very least, it does not seem the most worrying version of the problem. The conciliationist is motivated by the intuition that, in the canonical cases, lowering one's credence in the light of disagreement seems rationally required. It should not surprise the conciliationist that this applies to conciliationism itself. Solutions to the self-undermining problem that carve out a special exception for the theory of conciliationism, from conciliationism's own dictates, seem worrisomely ad hoc.

Elga (2010), for instance, carves out such an exception. Elga sees the need for responding to the objection that this "partially conciliatory" solution involves an ad hoc exception, and spends considerable space attempting to meet the challenge. His argument to this effect works by embracing enkratic principles like *most credence* and *weighted average*. Given these principles, any inductive method must carve out an exception for itself, on pain of suffering the same self-undermining inconsistency conciliationism faces. Thus, the move to partial-conciliationism is not ad hoc. It simply reflects the way every inductive method should be understood: as being inapplicable to itself. However, I think avoiding this worry altogether is a theoretical benefit. It seems like a cost to accept that all of our inductive methods are more complicated

than we realized. Moreover, we have good reason to deny the enkratic principles, given the benefits the theory of endorsement provides to inquiry. And this works to block self-undermining inconsistency for other inductive methods just as it does for conciliationism.

Conciliationism prescribes a certain method for updating beliefs (or, more carefully, is a general category of theories which prescribe such methods). Such methods are clearly not meant to be perfectly reliable at tracking the truth (or in increasing the accuracy of belief states). Indeed, it is precisely the fact that human subjects are error-prone that makes conciliationism plausible (if I was sure I was perfectly reliable, any disagreement with my peers in the mental math case wouldn't move me to want to conciliate). However, an imperfectly reliable method will sometimes lead to misleading evidence. There is no reason to think that this misleading evidence cannot be about the method itself. Thus, even if conciliationism is true, it should be no surprise that conciliating can (and does) offer misleading evidence in many cases, including in circumstances where the disagreement is about conciliationism.

Conciliationists, then, should expect that following the dictates of their own theory will sometimes lower their confidence in the theory. Any attempted solution to the self-undermining problem that denied this would be unsatisfactory, because it would fail to respect the spirit of the view. Thus, if any satisfactory solution to the problem will admit this, and since the view seems deeply committed to something like this, it cannot be that this is the essence of the self-undermining problem.

Recall that the self-undermining problem is meant to be a problem which (rationally) forces the conciliationist to give up her own view. It is not merely meant as an additional reason in favor of steadfasting, but as a problem which forces the conciliationist to give up the view based on their own commitments. But the fact that following the dictates of conciliationism can cause one to lower credence in the conciliationism is something that is in keeping with what the conciliationist wants. So it cannot, *on its own*, serve to force the conciliationist from her own theory. From the conciliationist's perspective, the lowering of credence in conciliationism in the face of disagreement is a feature, not a bug.

In order to make the objection pressing against the conciliationist, by her own lights, the objection needs to show not just that disagreement can (or even in the actual circumstances does) cause a rational lowering of credence. The objection must show that the conciliationist is forced to *give up* her theory, to give up conciliationism. That is, the objection needs to show that adopting conciliationism is somehow unstable or inconsistent. When "adopting" is understood as believing, and "giving up" as suspension, the self-undermining problem is extremely pressing. At least, it is pressing given our actual circumstances of pervasive, strong disagreement within the epistemology of disagreement. It is this pressing objection that I have reconstructed above, and I think that it is this version which explains why conciliationists have been worried by the objection.

I have argued that treating "adopting" and "giving up" in terms of endorsement, and denying the extra enkratic commitments listed above, solves this problem. The endorsement solution allows the conciliationist to respect the spirit and motivations of conciliationism by allowing that disagreement can lower one's credence in conciliationism, while at the same time vindicating the idea that one can maintain one's commitment to conciliationism.

A second worry I want to address is that, even if my analysis in the preceding section is correct, there is a part of the initial self-undermining objection that remains. Specifically, this worry concerns whether the conciliationist can adopt their theory, while remaining a conciliator in everyday life. We can call this the "practical objection" to the endorsement solution.

The basic idea of the practical objection is that the endorsement solution only applies to the domain of inquiry. It protects the rationality of committing to conciliationism during inquiry, but it fails to justify a commitment to conciliationism that is relevant outside of the epistemology of disagreement. That is, it fails to justify a practical commitment to conciliationism, in a way that leads to counter-intuitive consequences.

In order to address the practical objection, I am going to provide another version of Connie's case. This is meant to illustrate the purported problem, and the alleged

counter-intuitive consequences. I will then argue that this practical objection to the endorsement solution fails, because it does not actually lead to counter-intuitive consequences. First, the case:

> **Practical Mental Math Connie:** Connie proceeds as before, endorsing conciliationism, even after lowering her credence in it when faced with disagreement. Now Connie is at a restaurant with her friend, and they need to split the check. Connie calculates half the check to be $43, while her friend (who has the same evidence and the same level of competence at elementary arithmetic) calculates it to be $45.

What should Connie do, faced with this instance of disagreement? She endorses conciliationism, but she has a low credence in it, where "low credence" means the same as it does above: less than is required for justified belief. How should she now respond to this objection? A proponent of the practical objection suggests that, intuitively, she should conciliate. More importantly, they might say, any solution to the self-undermining problem should protect Connie's ability to rationally conciliate in these circumstances. But the endorsement solution to self-undermining doesn't provide justification for adhering to conciliationism here. Connie's endorsement of conciliationism rationally governs her behavior within the domain of inquiry: while she is arguing in the philosophy conference, seminar, or in a journal article. By the lights of the theory of endorsement, it does not (and should not) govern her behavior outside of this domain. Instead, it is her credences which should determine her behavior in practical situations of uncertainty, and according to my solution, her credence in conciliationism is low; perhaps even less than .5.

Thus, the practical objection alleges, there is something missing from the endorsement solution. It allows the researcher to be a committed advocate of conciliationism, but it does not protect conciliationism from being rendered un-followable in the practical domain. The endorsement solution doesn't solve the whole problem.[30]

---

[30]Thanks to Andy Egan, David Black, and Pamela Robinson for pressing this objection.

The response involves returning once again to considering which enkratic principles are plausible. Within research domains, I have argued that endorsement provides the appropriate normative framework, and this includes determining the correct enkratic principles. Outside of research, however, other principles must be provided. Here is one that seems highly plausible to me:

Justified Theory Principle Outside of research domains, if a subject has a justified belief in a theory, then the subject should follow any rules prescribed by that theory.

This principle is essentially the contrapositive of the Unjustified Theory Principle above, except that it only constrains subjects when they are outside of research domains. It is thus compatible with the endorsement principle.

Notice, however, that this principle does not constrain Connie, as long as her confidence in conciliationism has not been forced so low that it is compatible with a justified categorical belief in steadfastism (and so long as she does not meet whatever other requirements there are on justified belief). In Connie's practical mental math case, she has a categorical belief in neither conciliationism nor steadfastism, so the *Justified Theory Principle* does not apply. Connie has no relevant justified belief in a theory, and so the antecedent of the principle is not satisfied.

Furthermore, Connie has significant first-order, situation specific reasons to lower her confidence in light of disagreement in this case: the same reasons that one cites in explaining the intuitions that motivate conciliationism in the first place. That is, the disagreement shows that either Connie or her friend has made a mistake, and Connie has no reason to think it is more likely to be her than her friend. Since she is not confident enough in any general principle about disagreement to be constrained to follow its dictates in general, and since she has significant reasons to conciliate that are particular to this case, she should conciliate. Thus, the practical objection is blocked: a conciliationist does not need a solution to the self-undermining problem in order to be able to conciliate in specific cases where that is, intuitively, what she should do. The self-undermining problem only threatens her ability to be committed to the view

as a general principle.[31]

One might still worry, on this response, that one cannot be practically committed to conciliationism. But this is what the conciliationist should expect: after all, we conciliationists think that disagreement should lower our confidence in theories, and there is too much disagreement about conciliationism for it to be reasonable for us to apply it, across the board, as a rule for practical decisions outside of research domain. Hopefully, someday we will reach consensus in the field, and then we will be able to so apply it. In the meantime, the self-undermining objection offers no reason not to conciliate in the cases where it is clearly intuitive to do so, such as Connie's.

3.6 Conclusion

Conciliationism is a highly plausible theory in the epistemology of disagreement. However, its proponents face two problems which are meant to force us to give it up: the skeptical results and self-undermining problems. The theory of endorsement, however, offers us a way out of these problems. Endorsement is the appropriate attitude to take toward the theory of conciliationism during inquiry. The theory of endorsement rationalizes the committed advocacy of researchers toward their views, even after disagreement requires that they lose confidence in them. It also justifies distinct enkratic principles, and justifies the denial of some that have been taken for granted in the literature. This explains how it can be rational to hold philosophical and scientific views in the face of disagreement, and it explains how one can stably and consistently maintain a commitment to conciliationism itself. Moreover, the success of these solutions provides further evidence for endorsement, and its accompanying norms of inclusive epistemic rationality.

So, if you want to be a conciliationist, you should be an endorser.

---

[31] Compare this to Lasonen-Aarnio's appeal to particularism about higher-order defeat in (2013; 2014). If one thinks that no enkratic principles linking theories and rules are plausible, then the fact that Connie is not justified in believing conciliationism will have no effect on what she should do in a particular circumstances. If this is right, then there is no reason she should not conciliate whenever it seems intuitively appropriate to do so. However, it does seem plausible to many epistemologists that there are enkratic principles of this kind (which is why the self-undermining argument appears intuitively plausible in the first place).

Chapter 4

A Fragmented Solution to the Problem of Old Evidence

One of the most pressing problems for Bayesian Confirmation Theory (BCT)[1] is the *problem of old evidence* (POE). This problem, at its base, is that the standard ways of using Bayesian modeling entails that something one already knows cannot serve as evidence for any theory, new or old. But there are many instances in the history of science, and even in everyday inquiry, of thinkers considering how old evidence (things they already know) support, confirm, or provide evidence for a theory. Since BCT purports to be a "comprehensive and unified" model of inductive reasoning and confirmation, this is a significant problem. The problem is particularly unsettling because this is precisely what Bayesianism was supposed to do best.[2]

In this paper, I appeal to the growing literature on "fragmentation" to provide a solution to the problem of old evidence for Bayesianism. Appealing to a fragmented framework, which is independently well-motivated, provides a simple solution to this problem.

My purpose here is twofold. First, I want to show that fragmentation provides the best solution to the problem of old evidence for BCT. Second, I want to show that fragmentation has yet another application; one which is suggestive about a further set of applications involving limited information access by design. This provides additional evidence for the value of the fragmented framework for philosophy of mind and epistemology.

The first section of this paper provides a quick overview of the problem of old

---

[1] For more on BCT and its standing in contemporary philosophy of science, see Strevens (2006) and Earman (1992).

[2] For more on this thought, see Glymour (1980) and Garber (1983).

evidence. The second section explains fragmentation, and describes its independent support. Then, in section three, I provide the fragmented solution to the problem of old evidence. The final section details the advantages this fragmented solution has over the other solutions that have been offered for the problem.

4.1 Bayesian Confirmation and the Problem of Old Evidence

4.1.1 Bayesian Confirmation Theory

Bayesianism comes in many varieties. This is because there are many ways to add to the basic commitments of the Bayesian framework. Here, I will focus on a simple version of BCT which is characterized by three basic commitments. The first is personalism: that confirmation, inductive reasoning, and evidence should be based in part on modeling the attitudes that inquirers (usually scientists) have(and should have). That is, what is modeled is the rational update of a scientist's credences or degrees of belief as they learn new things. Thus, the norms of BCT are norms of rational belief update.

Second, these credences should be modeled by appeal to the probability calculus. That is, a rational credence function should obey the probability axioms, and be representable by a probability function. The subject's credence or degree of belief in some proposition H can be represented by a probability function that assigns a real number value between zero and one to H.

Third, that learning should happen in a way that can be modeled as update by conditionalization. That is, a subject's new credence, after learning something, should depend on her old conditional credences given that thing. These conditional credences are understood using the standard ratio definition: $Pr(H|E) = \frac{Pr(H\&E)}{Pr(E)}$. Update by conditionalization then requires that once the subject learns E, her new credence in H should be equal to her old credence in $Pr(H|E)$. Importantly for our purposes, when a subject learns E her new credence in E should be $Pr(E) = 1$. As we will see, it is this feature of BCT which led Glymour to recognize the Problem of Old Evidence (1980).[3]

[3]It is not essential to the Problem of Old evidence that evidence propositions receive credence 1 when

This basic BCT framework provides a variety of benefits. Importantly, BCT is conducive to a simple and intuitive definition of the *evidential support* relation in terms of probabilistic relevance. A is evidence for B (for a subject) just when $Pr(B|A) > Pr(B)$. Evidential support is explained in terms of probability raising. This use of conditional probability combined with update by conditionalization as a way of modeling confirmation has lead to the great successes of BCT.[4]

4.1.2 The Problem of Old Evidence

Update by conditionalization is at the core of BCT, and while it is responsible for much of its success, it also leads directly to the problem of old evidence. The POE arises because the evidential import of some piece of evidence is fully accounted for by the update procedure. After the evidence is learned, and after the update occurs, the proposition learned will no longer satisfy the probabilistic definition of evidential support.

The canonical example used in the literature to describe the problem is the evidential support provided for the General Theory of Relativity (GTR) by the orbit of Mercury (OM). Mercury's orbit was long known to be different than Newtonian mechanics predicted (Earman 1992, 119). In order to explain this well-known anomaly regarding Mercury's Orbit, Einstein developed GTR. Earman points out that, although general relativity has successfully made a number of further true predictions which confirmed it (gravitational lensing and gravitational waves for instance), apparently most physicists came to accept it before this evidence was available (1992, 119). Also, historical surveys of scientists suggest they took GTR's successful account of OM to be the most important factor in its confirmation (ibid).

---

learned. The POE arises even in frameworks that use Jeffrey conditioning and refuse to assign probability 1 to any contingent proposition. See Barnes 1999, sec. 6.2 for discussion of the "quantitative" POE (Barnes' own solution is to give up on the notion of evidence being propositional, which is an interesting move. Hopefully, the fragmented solution obviates the need for such drastic measures). For Jeffrey Conditioning see Jeffrey (1990). However, I will largely focus on the strict conditionalization framework, and the credence 1 version of the problem, for ease of exposition. The fragmented solution works just as well in both frameworks.

[4]For a detailed introduction to Bayesian epistemology and BCT, see Talbott (2015), Crupi (2015), Earman (1992), and Strevens (2006).

Given the facts of this case, it is clear that the orbit of mercury is evidence for GTR. However, BCT does not deliver this result. Einstein already knew about the orbit of Mercury (it's part of the causal explanation for why he formulated GTR). So at the time he formulated GTR, his $Pr(OM) = 1$, and therefore $Pr(GTR|OM) = Pr(GTR)$. Thus, by the probabilistic relevance definition of evidence, BCT falsely predicts that OM is not evidence for GTR.

Following Eells (1985) and Sprenger (2015), we can distinguish between two different versions of POE that might arise for a theory T and evidence E:

Static Version The subject learned E and T previously, so (the subject's) $Pr(E) = 1$, so $Pr(T|E) = Pr(T)$. But it is now sometime later and they wish to evaluate E's support for T.

Dynamic Version: E was known before the formulation of T, and it is now the time of the formulation of T (or barely after). Now the subject wishes to evaluate E's support for T.[5]

Einstein and his contemporaries faced the dynamic version of the problem. Einstein knew about the orbit of mercury before formulating the general theory of relativity. An example of the static version can be generated for anyone who already knows about both GTR and OM, but subsequently wonders whether OM really provides good evidential support for GTR.

Thus, there are two central types of cases concerning evidential support which BCT cannot account for. Bayesianism therefore makes bad predictions.

These two problems are significant for Bayesianism. Not only do we have the problem that something that once was evidence no longer is (static version), there are cases where things which intuitively should count as evidence never do (dynamic version). That is, according to BCT, if E is some proposition I already know, it simply cannot ever be evidence for any new theory. This looks bad. Bayesianism needs an interpretation that allows it to account for evidential support relations that exist atemporally,

[5]I have followed Sprenger's formulation of this distinction. Eells makes a few other distinctions that are useful and important, but do not concern us here.

while still being able to model learning over time. The simple, standard version of BCT cannot do both of these things at the same time.

What is at stake here is whether BCT is an appropriate model for inductive reasoning. The problem is that we want to use this simple mathematical modeling framework to model actual subjects' evaluations of evidential support, as well as their rational doxastic evolutions, and to derive norms that govern these things.

What has gone wrong, I want to suggest, is not that we are using the wrong mathematical tools. Instead, it's that the idealizing assumptions being made in order to apply these tools are incorrect. If we are going to use BCT to model creatures like us, we need to loosen or change some of the assumptions in order to get the right predictions.

The solution is to stop thinking that the model should be applied to a subject globally. Instead, BCT should be fragmented: the model should be applied to individual compartments of a subject's doxastic state. Once we do this, the model will apply without making bad predictions.

4.2 Fragmentation

Fragmentation is the idea that we should not model subjects as having one unified, global belief state. Subjects have access to different information at different times. Some information is accessible for one task, but not accessible for another task. We can represent this by treating the subject as if they are compartmentalized, or fragmented. That is, distinct parts of the subject's mind contain different information, and are operative for different purposes.

4.2.1 Motivations for Fragmentation

Fragmentation is intuitively plausible for creatures like ourselves, who have limited computational abilities and imperfectly reliable learning methods. It also does a good job of explaining a variety of difficult cases. I will touch briefly on a few of the main motivations for fragmentation. My purpose here is to illustrate the independence of

these motivations, as this is part of what makes fragmentation a compelling solution to the POE. It is not an ad hoc solution, but rather arises organically from a well-motivated and successful research program.

The notion of fragmentation first arose in the context of solving logical omniscience, consistency, and closure problems for possible worlds semantics. Lewis (1982) and R. Stalnaker (1984, 1999) both recognized these problems for a semantics where the meaning of a proposition is a set of possible worlds. Such a view has two features which lead to these problems: first, if a subject's information or doxastic state is represented by a set of possible worlds, this state must be both consistent and closed under deduction. Second, there is only one necessary proposition: the set of all possible worlds.[6]

It turns out that because of the requirements of the probability axioms, Bayesian epistemology has similar commitments. It requires that the subject have credence 1 ($Pr(T) = 1$) in any logical truth, and that she have equal credence in any logically equivalent propositions. This leads to Bayesianism having the same logical omniscience, consistency, and closure requirements as possible worlds semantics.[7]

There are a wide variety of extremely common cases which cause problems for these deductive requirements of omniscience, consistency and closure. I will just provide a few examples. Consider, for instance, Lewis's understanding of Princeton geography (1982):

> I speak from experience as the repository of a mildly inconsistent corpus. I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel... (436).

This, clearly, is an inconsistent triple. But Lewis's behavior, let's suppose, was not

[6]For more on possible worlds semantics, see R. Stalnaker (1984), Menzel (2016), J. C. King (2014).

[7]Strictly speaking, these requirements only hold for logical truths and equivalences which are logically expressed in the language the probability function is defined over. So the Bayesian is required to have $Pr(P \vee \neg P) = 1$ and $Pr(P \vee Q) = Pr(\neg(\neg P \& \neg Q))$. Still, this is enough to lead to major problems. And, if one wants to use possible worlds as the state space (the atoms of the language) that the $Pr(\cdot)$ is defined over, as many philosophers do, then the Bayesian is on the hook for all logical truths once again.

erratic or unpredictable. For some purposes, he behaved as if the street and the railroad were parallel (perhaps while drawing maps, or while giving directions) while for others he believed them perpendicular (perhaps when navigating himself to Nassau street from the train station).

Lewis appealed to fragmentation to explain his own case: his inconsistent beliefs were kept quarantined, as parts of different psychological compartments or fragments which were active for different purposes. His behavior in different circumstances is rationalized by different fragments. This does a better job of explaining the systematic character of the mistake, rather than simply pointing out that it is a mistake.

There are also failures of closure that are well explained by fragmentation. Suppose Julie has just finished painting a square room, and now needs to lay carpet.[8] She knows the room is square, and she knows how long the sides are: 9 feet. Moreover, she has a college education and knows that $9 \times 9 = 81$, and if you asked her how to calculate the area of a square she could tell you. However, when she is at the store buying carpet she realizes she doesn't know how much to buy.

Julie's case is not uncommon, and is certainly not impossible. But it is a clear failure of closure. The fragmentation framework has a simple way of explaining what is going on here: Julie has multiple doxastic fragments, and the fragment with the information about how to calculate the area of a square surface is not being accessed when she is in the store buying carpet. This explanation can be expanded to account for both failures of logical omniscience, and to explain mathematical achievement (Rayo 2013).

On top of the motivations for fragmentation that arise from these technical problems for possible worlds semantics and Bayesianism, there are a variety of other cases that are well explained by the framework. Some of the most compelling of these involve cases which seem endemic to the human condition: recall failures. Suppose I get asked the following question: (1) "What is a country in the world whose name contains three 'u's?" I might be stumped by this question, and not be able to answer. But suppose someone asks me: (2) "Is 'Uruguay' the name of a country that contains three

[8]This example is like the farmer example from Rayo (2013).

'u's?", to which I readily assent. How should we model my doxastic state? There is a sense in which I don't know that "'Uruguay' is the name of a country with three 'u's, because I couldn't answer question (1). But there is a clear sense in which I do know this proposition; the second phrasing of this question, in (2), elicits my knowledge.

Such examples are ubiquitous.[9] We often find ourselves unable to recall information which we know. And it turns out this is hard to explain if we are using a global model of a subject's doxastic states. Do I know that "Uruguay is the name of a country with three 'u's"? Which proposition gets placed in a representation of my belief state? What credence do I have in this claim? There is no obvious way to model my doxastic state here using a global set of beliefs or global credence function.

Fragmentation offers an easy solution for this as well. We should model subjects as having different belief states under different "elicitation conditions" or tasks (Egan 2008b, Elga and Rayo 2015). When prompted with the first question, I don't have access to the information about Uruguay. But when prompted with the second question, I do. This can be modeled by having two sets of beliefs, and/or two credence functions that govern my behavior under the two different elicitation conditions. When undertaking the task of answering the first question, I don't have access to the information about Uruguay, but when answering the second question a different fragment is operative and I do have access.

There are many other kinds of examples which admit of the same fragmentation treatment. Fragmentation can explain how implicit bias functions: one fragment contains beliefs with the bias, while another explains the explicit commitments which contradict this (Elga and Rayo 2015). Fragmentation also provides an account of cases where "knowing that" and "knowing how" come apart, and fail to interact. For example, it's well-known that one can be an excellent athlete, but make false claims about how one goes about exercising one's athletic competences. The fragmentationist can account for this by proposing one fragment for the "know how" the subject displays, and another fragment which governs the subject's (perhaps false) assertions regarding

[9] Cf. Egan (2008b), Elga and Rayo (2015), Rayo (2013) for more examples along these lines.

their own performances. Moreover, these examples fail to exhaust the framework's usefulness.[10]

In addition to providing an account of a variety of cases, fragmentation is simply intuitively plausible. It is a truism that we are limited creatures, with limited computational abilities. We don't have the capability of holding all our information in our working memory at a single time.[11] Given these limitations, ideals of rationality that appeal to global doxastic states seem overly idealized, at least for many purposes. Moreover, models of our cognition which appeal to global doxastic states seem unlikely to be accurate. Fragmentation offers a way of providing more achievable normative goals, and of modeling subjects more realistically.

It's important to note, however, that fragmentation is not motivated merely by the fact of computational limitations. There can be some cases in which even ideally rational agents will be better off being fragmented. Egan (2008b) provides examples of this kind of beneficial fragmentation. He suggests subjects have fallible "Spinozan" methods of belief-formation. A Spinozan method's operation directly results in a belief, without that belief being subjected to evaluation before being accepted. Perceptual beliefs are plausibly formed by Spinozan methods: when I see a cup, I immediately form a belief that there is a cup. I don't first weigh the evidence for and against this claim, and then reason my way to the belief. In cases where Spinozan methods get things wrong, it is better for an agent to be fragmented, so the errors don't seep into her entire doxastic state. Instead, they are quarantined within a few fragments, which can be corrected later.[12]

The fragmentation solution to POE will provide another instance in which even ideally rational agents will be better off if they are fragmented. The solution relies on fragmentation for the evaluation (and re-evaluation) of theories in the light of previously known evidence. This fragmentation is one of choice or design, rather than

---

[10] For more applications of the fragmented framework, see Egan (2008b), Rayo (2013), Elga and Rayo (2015), D. Greco (2014) , Yalcin (2015), and Johnson (2016).

[11] See Cherniak (1983) and J. Weisberg (2016) for support of this claim, though it hardly seems necessary to provide empirical support for such an obvious fact.

[12] For details, see Egan (2008b).

simply a way of dealing with computational limitations. And this will point the way to additional cases of fortunate and intentional fragmentation, as I will suggest in the concluding remarks.

4.2.2 How Fragmentation Accounts Work

Fragmentation is a framework in which we represent agents as being compartmentalized: their belief state is represented using more than one probability function.[13] Fragmentation just commits us to representing agents as having more than one discrete belief set, whether that be understood as a set of mentally represented sentences, or a set of dispositions to act, or even types of brain states.

In order to apply the fragmented framework to a Bayesian account, I will follow the standard practice of using indexed probability functions. Each subject has multiple fragments, each of which is represented by a probability function. Each probability function is indexed to indicate which fragment it is associated with. So instead of representing the subject's doxastic state with a single, global probability function, we instead represent the doxastic state with multiple probability functions, each of which is indexed and represents an individual fragment's associated degrees of belief.

I will follow (2013) and index in terms of tasks.[14] In particular, these will be tasks of evaluating individual theories. I will use a simplified version of Rayo's framework, since the POE is actually easier to solve than, for instance, the problem of logical omniscience.

A fragment is represented by a triple containing a task, a probability function, and a set of available background knowledge: $< t, Pr_t(\cdot), K >$. The t represents the task the fragment governs. The probability function will be indexed to this task: $Pr_t(\cdot)$. Thus, $Pr_t(P) = .5$ represents that the subject is .5 confident in P for the purpose of

[13] The framework is compatible with all of the leading views about the nature of mental states, e.g., representationalism, functionalism, dispositionalism, and radical interpretation. See Schwitzgebel (2015) for a summary of these various accounts. The only view it is clearly incompatible with is eliminativism.

[14] The fragmented solution is compatible with a variety of ways to index. See Egan (2008b), Elga and Rayo (2015), Yalcin (2015), and Johnson (2016).

performing task t. K represents the accessible background knowledge of the subject, accessible from the relevant fragment.[15]

The use of K to represent a subject's background evidence is standard in a Bayesian framework. A subject's probability function at a time is always conditional on their background evidence at that time: $Pr(P|K)$. The only difference here is that each of a subject's fragments can have a distinct set of background evidence K. In most Bayesian literature, this K is assumed and then suppressed, so that it is understood that a probability function written as $Pr(P)$ is equivalent to $Pr(P|K)$.[16] The composition of K, however, is important for the fragmented solution to POE, so I will continue to make this explicit. What counts as accessible background knowledge for a fragment, and so what will be in K, will depend on the fragment and the subject in question. I will return to the characterization of K, and how it should be constrained, in section 4.3.2 below.

This kind of framework can easily represent what is going on in the motivating cases from the previous section. For instance, we can account for the "Uruguay" recall case by appeal to two different indexed probability functions. When I am tasked with answering question (1), the version of the question that doesn't explicitly mention Uruguay, I am well represented by a probability function where I assign low credence to the claim (U) that Uruguay is a country whose name has three 'u's, so e.g. $Pr_1(U) = .01$. This is because I will behave, in answering the question, as though I have very low confidence that "Uruguay" is the right answer. However, when performing the task of answering question (2), where Uruguay is explicitly named, I am well represented by a probability function that assigns very high probability to U, e.g. $Pr_2(U) > .99$. This explains the systematic way in which my recall failure occurs regarding Uruguay and the spelling of "Uruguay": different fragments are active for different tasks, and these fragments have access to different information, which is represented both by the prior

[15]This is a slight departure from Rayo's (2013) use of this kind of notation, but the differences are unimportant here.

[16]For more on this usage, see Earman (1992). Note that K is not the name of a set, but rather a variable which takes a set as its value.

probability function of the fragment, and the membership of K.[17]

With this fragmentation framework on the table, I will turn to showing how it can solve the problem of old evidence.

4.3 The Fragmented Solution

4.3.1 The Solution

The fragmented solution to the problem of old evidence is to model the task of evaluating a theory using its own dedicated fragment. The basic idea is that we treat each evaluation (or re-evaluation) of a theory in light of available evidence as its own task. There will be a probability function associated with each such task, and a set of relevant background knowledge. Whenever a subject evaluates a theory, this involves a separate fragment. In this separate fragment, the previously known evidence need not be represented as having probability 1. I will first present this schematically, then apply it to the canonical example of general relativity.

Suppose a subject is deliberating in order to evaluate a theory T. She already knows E. However, she wants to consider E's import for T. So, when turning to the task of evaluating T, she has (or constructs) a fragment which we represent with $< t, Pr_t, K >$. Since she wants to evaluate E's import for T, she does not include E in K. This is despite the fact that in other fragments (e.g., for the task of answering questions about scientific knowledge, call it "*ask*"), her credence is $Pr_{ask}(E) = 1$. In the new fragment, $Pr_t(E) < 1$. But there is still a sense in which the subject continues to possess the evidence: because she maintains credence 1 in the evidence in *ask*. The fragmentation framework has opened up the space to represent her evaluation of the new theory as a new fragment, one which does not assign credence 1 to the evidence.

This means that we can represent the subject as having E be evidence for T. We can do this in the traditional Bayesian way, using the standard definition of evidential

[17] Of course, I won't be in a position to report my low probability in U for task 1, which makes the view incompatible with a certain kind of accessibilist internalism about a subject's own credence function. But this actually makes sense, given the commitments of the framework: if I am asked to report my credence in U, that is a different task. And plausibly, if you mention U, I will no longer have a low credence in it, for the same reasons that I have a high credence in U when asked question 2.

support. So if E really is evidence for T (for this subject, given her priors and accessible background knowledge), then $Pr_t(T|E\&K) > Pr_t(T|K)$. This leaves BCT to function in its usual way. The standard Bayesian framework is preserved, but relativized to individual fragments in a subject's mental economy.

In any fragment that involves evaluating the import of old evidence on a theory, K will contain all of the appropriately accessible information, but will not contain the evidence being evaluated. This is what preserves the evidential relationships. In section 4.3.2, I will discuss how both the subject's priors, and their K, should be constrained in different fragments. The basic idea, however, is that K should contain all the background knowledge which is genuinely relevant for the task, except for the E to be evaluated.

In addition to determining whether some piece of old evidence counts as evidence for a theory, we may also be interested in modeling the subject's considered judgment about the theory after considering this old evidence. This can be modeled in the standard way, as the subject's posterior credence (but now limited to within a fragment). After the subject considers the evidence to be evaluated, E, it can be treated as though it was learned. That is, the fragment is updated by conditionalization. Thus, the posterior probability the subject has (within the fragment) will be given by $Pr_t(T|E\&K)$. This represents the results of the subject's deliberative evaluation of E's import for T.

Note that this K does not represent a set of background information that is limited only by the computational limitations of human subjects. Instead, it is being intentionally limited, for the purposes of evaluating a theory. This is another example (like Egan's 2008b) of fragmentation that is good even for ideally rational agents. This intentional limitation can be thought of as a necessary, rational part of giving a fair evaluation (or re-evaluation) of the theory. K is limited to what is appropriately accessible, not simply to what is accessible given the abilities of the subject. Contrast this with the imperfect recall cases from section 2.

This fragmented solution can explain eboth types of old evidence problem cases identified in section 4.1.2. As we will see, this gives it an immediate advantage over

other proposed solutions. The solution can be illustrated by variations on the canonical case involving Einstein and the general theory of relativity.

First, I will consider the dynamic version, the historical version that reflects Einstein's own situation. This version of the problem is where the evidence was known before the formulation of the theory, but it is now the time of the formulation of the theory (or barely thereafter). Recall that Einstein knew about the anomaly presented by the orbit of Mercury (OM) prior to Einstein's formulation of the General Theory of Relativity (GTR). Nonetheless, Einstein, his contemporaries, and subsequent physicists have taken GTR's explanation of the orbit of Mercury to provide significant evidence for relativity.

The fragmented framework explains this case as follows. For each individual who evaluates GTR in light of OM, and OM's evidential import for GTR, we can represent this task of evaluation by a separate fragment of their doxastic state. Consider Einstein himself: when he first formulated and understood GTR, he already knew OM. However, when undertaking the task of evaluating GTR in the light of the evidence, we represent him as opening a new fragment of his doxastic state, indexed to this task. Let's call the task "evaluating relativity," or er.

The relevant fragment of Einstein's doxastic state is then represented by $< er, Pr_{er}, K >$. K represents the background knowledge that is both relevant to the task and computationally accessible to Einstein on the occasion(s) of evaluation. Selecting the right K depends on substantive epistemic and psychological considerations. This selection is not easily operationalized, but is nonetheless of a kind which scientists and philosophers make routinely (more on this below).

Given what we know of the case, when K is appropriately selected this should provide us with a fragment in which Einstein's $Pr_{er}(OM|K) < 1$, and his $Pr_{er}(GTR|OM\&K) > Pr_{er}(GTR|K)$. Thus according to BCT's standard definition, the orbit of Mercury was evidence for the General Theory of Relativity. Meanwhile, in other fragments, such as the one indexed to the task of teaching astrophysics, Einstein maintains high credence in OM, and thereby continues to possess the evidence.

Thus, the fragmented solution allows BCT to make the right predictions in the

general relativity case, and for the dynamic version of the POE more generally.

The fragmented solution is essentially the same for the static version of the problem of old evidence. In this case, we again simply represent the subject confronted by the problem as having a new fragment, one indexed to the task of evaluating the theory at that time.

As an example, consider a contemporary physicist. She already knows about OM and GTR. If she stops to (re-)consider what kind of evidence OM provides for GTR, we can represent this consideration as a task with its own associated fragment of her doxastic state. This fragment represents the task of evaluating relativity for her in 2018, so we might call it $er2018$. This fragment will have an associated probability function, $Pr_{er2018}(\cdot)$. From there, the example can be handled precisely the way we handled the above case for the dynamic version.

So, the fragmented version of BCT provides the right predictions in both versions of the Problem of Old Evidence.

To recap, the solution to the POE is to model the evaluation of a theory as a separate fragment in a subject's doxastic state. This is meant to represent the act of evaluating as a separate cognitive task. As in any Bayesian story, this doesn't require that the subject actually calculate her probabilities, or think explicitly about what her available background knowledge is. The model of a fragment is instead meant to represent an agent's act of deliberation or evaluation, and to make predictions about their future behavior based on this.

4.3.2 Constraints on Fragmentation

The biggest worry for the fragmentation framework in general is how to constrain the fragmentation. There are two parts to this. First, how to constrain how many fragments we use in representing the subject's behavior. This is a significant worry, because it might be tempting to simply add a fragment for each behavior the subject engages in. But this would leave the framework rationalizing almost any behavior, and it would offer little predictive power. However, this part of the worry for fragmentation is not pressing for the current solution on offer. Instances of freshly and

fairly evaluating theories with an open mind, in light of accessible evidence, seem like familiar occurrences. It's plausible that we can recognize such cases, and they form a reasonably natural kind. So indexing to these particular tasks seems plausible enough.

The second part of the constraint issue involves constraints on how different fragments relate to one another. There are conflicting considerations here, pulling us in different directions. On the one hand, the framework is designed to deal with inconsistency and recall failure. Thus, fragments must be able to differ significantly from one another. On the other hand, if there are no constraints on what the new fragments look like, fragmented BCT might still not give us much in the way of predictive and explanatory power, and may just appear too permissive.

For the fragmented solution to the POE, there are two things that must be constrained: 1) the prior probability for the function of the new fragment, and 2) the set K.

The notion of "prior probability" that must be constrained here is sometimes called the *ur-prior*. In the traditional Bayesian picture, this is the (perhaps hypothetical) probability function describing the subject before they learned anything at all. The subject's degrees of belief at a time are given by conditionalizing the ur-prior on all evidence and background knowledge the subject possesses at that time. On the fragmented picture, each fragment will, formally speaking, have it's own ur-prior. This is the prior probability before conditionalizing on K, the relevant set of background evidence.[18]

The first question of how to constrain fragment similarity, then, is a question about how the ur-prior of a new fragment has to relate to the ur-prior of the subject's other fragments. There are a variety of reasonably plausible ways of answering this question. In order to provide a concrete solution to the problem of old evidence, I will endorse one such way: what we can call *subjective consistency*.

Subjective consistency requires that each individual subject has their own ur-prior, and that this ur-prior remains constant in all of their fragments. In keeping with

[18]Note that the fragmentationist, like the traditional Bayesian, is not committed to there ever having been a time that the subject's beliefs are accurately described by the ur-prior.

the tradition of subjective Bayesianism, we can allow that this prior be any regular probability function, that is, any probability function that only assigns probability 0 to logical contradictions.[19] Differences between a subject's fragments are then entirely explained by differences in the background knowledge K available in the fragment.

The subjective consistency requirement ensures that fragmented subjects do not count as rational if they have wildly disparate fragments. Moreover, it ensures that when we model subjects who are evaluating the evidential support provided by old evidence, they are not rationally permitted to skew their results in virtue of selecting a favorable ur-prior for the fragment. They are instead required to remain consistent with their own starting place.[20]

The second way that fragmentation of a subject might be rationally constrained concerns the membership of the set of background knowledge K in any fragment. This is important to the fragmented solution. In order for fragmentation to solve the problem of old evidence, the fragment describing the subject's process of evaluation needs to exclude the evidence to be evaluated from their background knowledge. That is, K must exclude E. But there are many different such sets. How should this new set K differ from the subject's other fragments, and how should it be the same?

I think that there will be no fully general answer to this question. What should be in K in the subject's new fragment will be all the of *relevant* background evidence. The most obvious way that relevance will be determined is by appeal to the kind of inquiry that the subject is engaged in. In the General Relativity case, for instance, what will determine which propositions get into K will be a combination of the norms and standards of physics, and the subject's competent judgment about which things are relevant.

---

[19] This is what Weisberg calls "Initial Regularity" (2011).

[20] Of course, there are other options for constraining ur-priors. An objective Bayesian might argue that everyone should have the same ur-prior in every fragment (Earman (1992), Strevens (2006), Talbott (2015)). There are also even more subjectivist options: perhaps a subject could have a variety of different ur-priors, as long as each one was close enough to the others, as judged by some distance measure like the Brier score (Joyce 1998; Pettigrew 2016). One could even argue for a fully unconstrained subjectivism: whatever seemed right to the subject at the time. Which view one prefers depends primarily on commitments beyond the simple BCT framework I have been presupposing here. Any of them are compatible with the fragmented solution. I chose subjective consistency as a plausible way of making the account concrete.

This informally characterized constraint does not entail radical permissivism: in each case, there may be a fully determinate answer to what counts as relevant background knowledge. The point is simply that this will not be something which can be adjudicated formally, in full generality, without appeal to the particular features of the case. In Williamson's terms, it is not "operationalizable" (Williamson 2008). I think this is what we should expect: what counts as relevant background knowledge for a task will depend on substantive epistemic and psychological considerations about the subject and the task at hand.

This leaves us with the following picture: the new evaluation fragment is constrained by the subject's particular ur-prior, and on by features of the subject particular task which determine which background knowledge counts as evidence. These relatively minimal constraints allow the solution to get the right results in POE problem cases, while avoiding excessive permissiveness.

4.4 Advantages over Other Solutions

The fragmented solution to the problem of old evidence is intuitively plausible, independently well-motivated, and gives the right verdicts in each kind of case. Moreover, it does all of this without significant deformation of the formal machinery of Bayesianism, and thus preserves BCT's advantages.

In the last section, I detailed how the solution gives the right verdicts in the problem cases. It is intuitively plausible because it represents the way subjects deliberatively evaluate the support provided for a hypothesis by a set of accessible evidence. The compartmentalization of the subject's doxastic state is not ad hoc, but tracks the kinds of tasks and activities subjects actually undertake. It is independently motivated because it is situated in the larger fragmentation project. It makes use of the same tools that are used to solve an array of problems (see section 2 above). It is not a solution merely targeted at the problem of old evidence. Rather, the framework was designed for other purposes, and is only now being applied to this project.

There is also some empirical evidence that speaks in favor of this solution. This

evidence is far from settled, but I think that it is suggestive. The evidence is that subjects in experimental settings are influenced in their level of confidence by the fluency with which they recall things (Weisberg 2016, Overschelde 2008). That is, the more difficulty a subject has in recalling something, the less confident they become that it is true. That is to say, the faster a subject can recall something, and the more they can recall about it, the more confident they become in their answers. This effect has been elicited through various interventions, including increasing the time the subjects must work on a task (Koriat 2006), how much they were required to recall from the task (Koriat 2008), or making the reading more difficult by covering some of the letters (Whittlesea et al. 1990). Weisberg has argued, I think persuasively, that this is evidence that subjects in fact construct their credence functions on the fly, in the situations they find themselves in, based on the ease of their recall.[21]

This empirical evidence for construction of credences on the fly is predicted by the fragmentation framework, and in particular by the application I am proposing: that scientists construct a new fragment when they are evaluating a new theory in the light of previously known evidence that they are now recalling. A full elaboration of this empirical argument for the fragmented solution would require additional empirical work, and is beyond the scope of this paper. Nonetheless, I think this evidence is suggestive.

So the solution is intuitive, independently motivated, makes good predictions, preserves the formal machinery of Bayesianism, and enjoys at least modest empirical support. What remains to be shown, however, is that it is the best solution. In this section, I will argue that it has significant advantages over the most prominent alternative solutions on offer.

---

[21] See Overschelde (2008) for a helpful overview of this empirical literature, and J. Weisberg (2016) for the application of this literature to philosophy. He uses it to argue for pluralism about belief and credence, a view which I also endorse. However, the argument has clear implications for the fragmented solution to POE, and fragmentation in general, as I briefly sketch below.

I will focus on the two most prominent proposed solutions to the POE.[22] I will argue that each of these solutions is in fact committed to a kind of fragmentation already. However, each of these solutions makes additional commitments which cause trouble for them. These additional commitments are unnecessary; all that is needed to solve the POE is fragmentation, as we have seen. Also, the explicit appeal to fragmentation, which embeds the fragmented solution into the wider, independently motivated fragmentation framework, gives a better justification for the use of fragmentation.

4.4.1 GJN

The first alternative solution is due to Garber (1983), Jeffrey (1983), and Niiniluoto (1983). This is sometimes called the GJN-style solution, or the Garber-style solution.

The key move in the GJN solution is to suggest that the subject learns something else when they consider how old evidence relates to a new theory. That is, when they apply the old evidence to the new theory, they are not updating directly on the evidence. Instead, they are learning a different proposition, one that is about the relation between the old evidence and the theory. On this view, the subject doesn't update her credence in T on E. Rather, they update on a sentence that says "E entails T". This gets represented by "$E \vdash T$", though here the "$\vdash$", can refer to any kind of logico-mathematical entailment, not just the consequence relation of propositional logic. According to GJN, then, when E is old evidence for T, what the subject is actually doing is learning that E entails T. Thus, it's possible that $Pr(T|T \vdash E) > Pr(T)$, which meets the definition of evidence.

However, this solution requires a way to relax logical omniscience requirements. Otherwise, the subject will be required to have probability 1 in $T \vdash E$, and that won't help at all. The GJN solution achieves this by interpreting sentences of the form $T \vdash E$

[22] The solutions I focus on are ones which do not require significant changes to the Bayesian formalism, and which don't give up on our three basic features of Bayesianism. Other solutions which require more radical changes to BCT (e.g., Wenmackers and Romeijn (2016)) thereby incur significant theoretical costs, though there is not space to consider the issue here.

as atoms of the language the probability function is defined over. This has two important results. First, that the subject need not have probability 1 in them. If such sentences were interpreted as entailments in the language, the probability axioms would require having full credence in them (since they are logical truths). It does not solve the POE to exchange one set of probability 1 sentences for another. So this involves a relaxing of logical omniscience, essentially by making the fact that "$\vdash$" means entailment opaque to the subject's probability function.[23] The second result is that any logical relationship between E, T, and $T \vdash E$ must be "extrasystematic," that is, a constraint that is added over and above the standard logical relations of the language.[24]

In order to solve the logical omniscience problem, the GJN solution requires that the state space, that is, the set of all possible states of the world over which probability is distributed, cannot be the set of possible worlds. This is because the states are determined by the truth values of the atomic sentences (there is one state for each consistent truth-value assignment to all atomic sentences). Since some of those sentences are actually logical truths in disguise (like $T \vdash E$), the state space must include states where such logical truths are false. But such *impossible* states are not among the *possible* worlds. If we take possible worlds as the state space of our probability function, entailment relations like $T \vdash E$ will be true in all of the worlds still in the subject's space of possibilities, because all necessary truths are true in every possible world. So it would be impossible to be uncertain about $T \vdash E$. And updating on certainties, as we have seen, is the problem.

The GJN solution thus tries to provide a solution to logical omniscience as a way

---

[23] In their version of this kind of solution, both Hartmann and Fitelson (2015) and Sprenger (2015) dispense with treating $\vdash$ as entailment, instead viewing it as "explanation." Their versions of the solution divorces the solution to POE from the problem of logical omniscience, which I think is the right move (though note the objections to this move in Kinney (2017)). However, the remaining objections to the GJN solution still apply.

[24] Even all these changes are not yet enough to allow the GJN proposal to solve the problem of old evidence: we also need some further extrasystematic constraints that ensure that the extrasystematic entailment claims actually bear the proper evidential support relations needed to make the right predictions in old evidence cases. There have been a profusion of suggestions for how this might work (Earman 1992; Hartmann & Fitelson 2015; Jeffrey 1983; Kinney 2017; Sprenger 2015). Most of the literature on the GJN solution focuses on giving an account of these extrasystematic constraints that gets the right results and is well-motivated. No version of these constraints has won consensus acclaim. I won't focus on that problem for GJN here, as I think the proposals problems run deeper than that.

of solving the POE. However, this solution to logical omniscience carries a number of worries with it. The first problem, recognized immediately by Garber (1983), is that it means the subject is required to be certain of arbitrarily complex tautologies expressed in the language the probability function is defined over, but allows them to be ignorant of extremely simple, even obvious logical truths. Supposing P, Q, and R are sentences in the relevant language, the subject would have to be certain of (the non-obvious tautology) $((\neg P \& P) \& Q) \rightarrow R$, while being allowed to be uncertain of $P \vdash P$.

In order to distinguish these two cases, and in order to make it seem less problematic that he is giving up on the atoms of the algebra being possible worlds, Garber appeals to a distinction between "local" and "global" Bayesianism. Garber suggests that the framework of BCT should not be something that applies globally to a subject, but is applied to a particular problem using a problem-relative language L. That is, for describing a particular scientist dealing with a particular problem, we only require "local" omniscience, relativized to the language best suited for the problem at hand. This allows the language to have uncertainties about non-empirical truths, like $T \vdash E$, by making them atoms of the language, while still using the Bayesian framework (which requires logical omniscience within the language).

This local/global distinction should sound familiar: though he doesn't use the term, Garber is essentially suggesting fragmentation. He does so to deal with these issues arising from logical omniscience. However, to solve the two issues above, he is also changing the language of each fragment, giving up on possible worlds, and describing the way scientists update on old evidence in a way that is, I will suggest below, counter-intuitive.

Thus, the GJN solution really is a fragmented solution. However, what its proponents have not realized is that the fragmentation is all that is necessary to solve the problem of old evidence. The extra features of the view involving changes to the language, additions of "extrasystematic" items and constraints, and treating logically complex statements as atoms, cause it unnecessary trouble. Moreover, the view tries to solve the POE by solving the more difficult problem of logical omniscience first, which is unnecessary.

Perhaps the most difficult problem the GJN solution faces is that it involves interpreting all old evidence cases as updating on propositions like $T \vdash E$. But it is plausible that actual scientific episodes involve cases where this doesn't seem to have been the case. In fact, Earman argues that Einstein took OM to be his evidence for GTR, not some fact about what is explained (1992). He suggests that OM and $GTR \vdash OM$ are "neither semantically nor extensionally equivalent". We have evidence that Einstein took OM to be his evidence, but no evidence that he took $GTR \vdash OM$ to be evidence. And regardless of how it worked for Einstein, it does seem like someone could possibly take the one thing as evidence and not the other. Moreover, Earman suggests that the fact that GTR entails OM must also be old evidence for contemporary students who want to consider OM's support of GTR. This is because the students almost always learn that GTR entails or explains OM before they really learn any details about the theory itself. So the GJN account claims that all old evidence cases are to be explained by appeal to updating on these entailment claims. But this is implausible in many cases.

This last case, involving physics students, raises another significant issue for GJN: it is only a solution to the dynamic version of the POE. GJN does not give the right prediction in first static version cases. Once a subject updates on a $T \vdash E$ claim, it will be assigned probability 1 and will no longer meet the probabilistic relevance definition of evidence. The GJN solution does not offer any way to make it that something remains evidence for a theory at any time after it has been learned.

The fragmented solution does not suffer from these problems, and has significant advantages. It is simpler, because all it requires is the fragmentation. It involves fewer commitments, and does not make us give up on possible world semantics. Meanwhile, it derives independent support because it is embedded in a larger fragmentation framework. Moreover, it provides a solution to both versions of the POE. Finally, the fragmented solution allows OM itself to be evidence for GTR even when it is old evidence, and has an explanation for how to make this very same old evidence "new" and relevant again. Thus, the fragmented solution is preferable to GJN.

4.4.2 Counterfactual Solution

The second alternative solution I will address is the counterfactual solution, from Howson (1991) and Howson and Urbach (2006). The basic idea behind this solution is that we can explain old evidence by appealing to a different probability function than the subject's current one. Specifically, this is the probability function the subject *would* have had, if they had not already learned the evidence. Since this counterfactual probability function would not treat the subject as having probability 1 in the old evidence, it would be able to have $Pr(T|E) > Pr(T)$, as is required for the definition of evidence. The trick is just to find the right counterfactual probability function. This would be a function that starts with the subject's own ur-prior, but is then conditional on a set of background knowledge containing all the subject's knowledge, but with the evidence to be evaluated, E, left out: $K - \{E\}$. Unfortunately, this trick turns out to be difficult.

Notice, this solution again has much in common with the fragmentation solution. It appeals to an additional probability function, distinct from the one that represents the subject's current belief state. Thus, there are two probability functions that describe the subject's doxastic state, rather than a single global one. Moreover, this function must be conditional on a set of background evidence K that does not contain E. So the solution is already committed to something like the fragmented solution. Formally speaking, they are quite similar.

The counterfactual solution has much intuitive appeal, but it suffers from significant problems, ones which again arise from the non-fragmentation aspects of the view. The first problem is that there might not be any probability function the subject would have had if she hadn't learned E. To take an extreme but illustrative example, it might be that learning E saved the subject's life, so that there are no nearby worlds in which the subject didn't learn E and has any doxastic state whatsoever.[25]

Second, even if there are probability functions the subject would have counterfactually, there is no guarantee that there is only one. That is, it is very unclear that there

[25] See Earman (1992) and Strevens (2006) for more on this line.

is a single probability function that satisfies the description "the probability function the subject would have had if she had not learned E." This is for a number of reasons, having to do with the logical relations between E and other members of K. These problems are related to the problems in the logic of belief revision literature regarding how contraction, replacement, and iterated belief updates work (Huber 2013a). The problem is that there are multiple ways to obtain a set of propositions closed under entailment that satisfy the constraint of not including E.[26]

The problem for the counterfactual solution is that these different ways of removing E to arrive at $K - \{E\}$ (and associated new probability function) might disagree with respect to whether E is evidence for T, and determining this is the point of the exercise.

So some cases will contain no such counterfactual probability function, and other cases will contain too many. But even if these problems were solved, perhaps by appeal to the best theory of the logic of belief revision (Huber (2013b)), there would still be the question of why what the subject *would have* believed is relevant. BCT is supposed to tell us whether something *is* evidence for a theory, not whether it would be. This seems to be a bit like changing the subject.[27] So the counterfactual solution, like GJN, threatens to involve a kind of change of subject: it tells us whether some proposition is evidence for another, if things had been different than they are. But we want our Bayesian framework to get the right answer about whether some subject has evidence for a theory, given the way the world actually is.

The counterfactual solution is committed to something like fragmentation. Fragmentation is all that is needed to solve the problem of old evidence. The additional commitments the counterfactual solution makes are about counterfactual relations

[26] To see why this is the case, consider a set K closed under entailment which contains a large number of sentences, including R. Suppose we remove R from K. We must also remove any sentences which entail R. However, there are some sentences which do not individually entail R, but do together. For instance, both $P \rightarrow R$ and P entail R. Which do we remove? Removing either will avoid the entailment of R. But it is unclear which removal we should prefer. Various solutions have been proposed for these problems, but none have gained consensus support. And the problem becomes much worse when such revisions are iterated. For more on this topic, see Gillies (2001), Levi (2004b), Huber (2013a, 2013b).

[27] See Earman (1992) and Strevens (2006) for similar worries.

between the subject and certain probability functions, relations which might not obtain to functions that might not exist. The counterfactual nature of these relations is what causes the problems, and this counterfactual relation is not needed to solve the problem of old evidence.

The simple fragmented solution I have been presenting is, from a formal standpoint, very similar to the counterfactual solution. Each requires that we identify some set of background knowledge K, and some probability function conditional on this K, which does not include the evidence E in question. However, where the solutions differ is in how this set is determined. Howson and Urbach appeal to a subtraction operation and a counterfactual analysis, whereas the fragmentation solution appeals to the independently motivated idea of fragmentation.[28]

Fragmentation thus presents a different answer about the nature of the subject's different probability functions and their relations to one another. Since the fragment utilized for the solution is not derived from other fragments via contraction, it doesn't suffer from the problem of being indeterminate between different functions. Moreover, the appeal to a separate probability function is justified not because it is what the subject would have believed. Instead, it is justified by its use in describing something everyone admits is going on in the cases in question: the subject's deliberation about whether some evidence supports some theory. This avoids the worry that there is no relevant probability function. Moreover, it avoids any worry about changing the subject. The fragmented solution uses a probability function that describes the actual state of the subject to explain why E is evidence for T. This function is just limited to a particular compartment of the subject's doxastic state.

Sprenger (2015) offers a solution which combines the counterfactual solution with the GJN solution. On Sprenger's account, the relevant counterfactual probability function describes what a hypothetical scientist, who is well informed but unaware of

[28]Note also that, although I am focusing on a version of the fragmented solution which is committed to subjective consistency of the subject's ur-prior, that is not the only option for implementing the solution (as noted in fn. 20). If we relax this constraint, the fragmented solution has another advantage over the counterfactual solution: it would be more permissive in allowing for changes in how the subject evaluates evidence, because it would allow for different ur-priors in different fragments.

E, *would* believe in the relevant scenario. This counterfactual scientist's probability function is augmented with GJN-style extrasystematic constraints. This solution also clearly relies on a kind of fragmentation, but ultimately suffers from both the problems I have raised for GJN and and for Howson and Urbach's counterfactual solution.[29]

The fragmented solution, therefore, has significant advantages over the counterfactual solution.

Conclusion

The problem of old evidence threatens Bayesian confirmation theory by suggesting that it fails to deliver on one of its purported best attributes: describing and explaining the evidential support relation. The fragmented solution provides a response to this worry that is independently motivated, intuitively plausible, and explanatorily fruitful. It is also compatible with possible world semantics. Although the fragmentation framework has been used in a solution to logical omniscience (Rayo 2013), the fragmented solution to POE does not rely on solving this much harder problem first. Nor does it require significant alterations to the Bayesian framework. Instead, it simply relaxes an assumption often left in the background, and a poorly justified one at that: the assumption that a single probability function should be applied globally to a subject in all circumstances. The fragmented solution preserves BCT by suggesting probability functions describe (and rationally govern) fragments of the subject's doxastic state.

The fragmented solution also enjoys significant benefits over its main rivals as a

[29]There is another related solution that involves distinct kinds of probability functions, first suggested by Eells (1985), and then expanded on by Eells and Fitelson (2000) and James Hawthorne (2005). This solution posits two distinct probability functions used to describe an agent: one representing the agent's evidential support judgments, and one representing their current belief state as it is updated over time. Whether *E* is evidence for *T* depends on the first function, while whether *E* confirmed *T* depends on the history of the second function. This solution is also clearly committed to a form of fragmentation. It would thus benefit from being embedded in the wider fragmentation framework. I think it is a cost that this theory splits up the probability function that describes evidential support from the one that determines confirmation, as one benefit of BCT is that it explains both of these in the same way. However, there isn't space here for a detailed treatment of that objection. What is worth noting is that this further alternative solution is also committed to fragmentation.

solution to the POE. In fact, these rivals are best understood as imperfect versions of the fragmented solution, which make unnecessary, problematic commitments.

Thus, I suggest that the fragmented solution does solve the Problem of Old Evidence.

Moreover, it points us in the direction of solutions to other problems. For instance, I conjecture that the problem of novel theories (Strevens (2006), Earman (1992)) admits to a similar treatment.

This application of fragmentation to the POE also expands the framework. It does this by treating limited accessibility to information not simply as something caused by lack of computational power. Instead, access to information can be limited by choice or design, even for ideally rational agents, in order to appropriately consider and re-consider evidential support relations. This notion could be widely applicable in describing what goes on in cases where evidence is intentionally restricted, e.g., in legal settings where certain evidence is excluded, or scientific settings where only evidence of a certain quality is admitted.

Chapter 5

Virtuous Distinctions: Virtues of Knowledge and Virtues of Inquiry

A version of this chapter is published in *Synthese* 194 (8):2973–3003 (2017).

5.1 Virtue Epistemology: A House Divided

Virtue epistemology is a family of epistemological theories which take some notion of virtue or competence as their central explanatory concept. This family has been divided into two camps. *Virtue reliabilism* uses the concept of a virtue or, synonymously, a *competence* to solve traditional problems in epistemology. Ernest Sosa (2007, 2010) and John Greco (2010), for instance, each offer an analysis of knowledge in terms of competences.[1] Competences are dispositions of subjects that serve as reliable methods of belief formation. Virtue reliabilism thus moves the locus of epistemological evaluation from exclusively focusing on belief states and propositions, to focusing on features of subjects and their performances (see Sosa 1980 and Battaly 2008).

Virtue responsibilism seeks to push the locus even further onto the subject. Responsibilists, such as Linda L. T. Zagzebski (1996a) and Jason Baehr (2011), suggest that epistemic evaluation should follow the model of Aristotelian virtue ethics. In virtue ethics, the primary bearers of moral value are character traits of subjects, *viz.* the virtues. If any states have value, or if there is any rule of right action, these things are ultimately dependent on the nature of the virtues. Correspondingly, responsibilists

[1] Greco seems to prefer the term "abilities," but we can set that aside for the purposes of this paper.

suggest that the primary bearers of epistemic value are epistemic virtues.[2] Moreover, responsibilists suggest that virtues are character traits for which we can hold the subject responsible (Axtell 1997; Montmarquet 1992). For this reason, they posit a distinction between intellectual character virtues, which are stable, person-level character traits of subjects, and those "virtues" that are mere cognitive faculties. Responsibilists argue that *virtue reliabilists* (and reliabilists generally) are mistaken in focusing primarily on cognitive faculties instead of person-level character virtues, since it is the character virtues that bear epistemic value.[3]

I will argue that there should be no reliabilist/responsibilist conflict among virtue epistemologists. The apparent conflict arises from the way the character virtue/cognitive faculty distinction has been drawn. I argue that this distinction is unhelpful; we should carve up the theoretical terrain differently. Once we recognize several other important distinctions among virtues, it will be clear that responsibilists and reliabilists are engaged in different projects, and that certain responsibilist critiques of reliabilism miss the mark.[4]

It is worth noting at the outset that I am not suggesting that there is no interesting distinction between the two *projects* that the reliabilist and the responsibilist are engaged in. On the contrary, I think these are two interesting and distinct projects worth pursuing. What I aim to show is that, once we make the right distinctions among virtues, we will see that these two projects are not in conflict. They are not trying explain the same things, nor make prescriptions about the same kinds of things. What I

---

[2]It's worth noting that many responsibilists, including Baehr and Zagzebski, also recognize the value of the truth of beliefs. However, we might distinguish this *alethic* value from epistemic values that go along with notions like *warrant* and *justification*.

[3]Though they differ on whether the faculties count as virtues, or are at all epistemically important. For instance, Montmarquet (1992) and L. T. Zagzebski (1996a) want to limit virtue talk entirely to character virtues, while Baehr (2011) and Battaly (2007, 2008) argue for the importance of such faculties in understanding some kinds of knowledge.

[4]I am not the first to suggest that the two projects are not in conflict, but are rather complementary (see Axtell 1997 and Battally 2007, 2008 for others making this kind of argument). However, other attempts to bridge the divide have relied heavily on the faculty/character distinction. Battaly, for instance, suggests that reliabilist faculty virtues can be used to explain "low-level" knowledge, and character virtues to explain "high-level" knowledge. I will instead suggest a different relationship exists between instances of knowledge and different kinds of virtues. My approach is thus entirely different, even if some of the goals are shared.

do want to replace is the "faculty/character" distinction. In particular, the notion of a "faculty virtue" should be abandoned, and we should understand character virtues in a different way.

I will also argue that, with better distinctions on the table, we can see that the virtue reliabilist project is in some ways more fundamental than the responsibilist project, since the latter importantly depends on the former. I will suggest that the distinctively *epistemic* value of responsibilist character virtues is dependent on their relationship with the competences studied by the reliabilists. This recognition of the dependence of the responsibilist project on the reliabilist one is not meant as a criticism of responsibilism. Rather, it is a way of securing the separate importance of each project by clarifying how they relate to one another.

5.2 A Responsibilist Challenge to Reliabilism

Responsibilist virtue epistemology is modeled on Aristotelian virtue ethics and focuses on global character traits of the subject. Call such a trait an *intellectual character virtue* (ICV): a person-level intellectual excellence of character. These are character traits for which it makes sense to hold the agent responsible for having, hence the term "responsibilist." Baehr defines an ICV as "a character trait that contributes to its possessor's personal intellectual worth on account of its involving a positive psychological orientation toward epistemic goals" (2011, 102). This latter notion is akin to personal *moral* worth. Similarly, Zagzebski requires that a virtue be "an acquired excellence of a person in a deep and lasting sense," one which is acquired by hard work over time, is not merely a skill, and is appropriately motivated (1996, 135).[5]

In order to contribute to personal intellectual worth, and for it to be a trait the

[5]Baehr's responsibilism is what he calls "weak conservative VE," and Battaly (2008, 643) calls "virtue-expansionism." The theory is *conservative* in that it has implications for traditional problems in epistemology, such as the nature of knowledge and the normativity of evidence. It is *weak* in that Baehr does not think that it can provide all the answers to traditional problems (e.g., he does not think that there is a plausible analysis of knowledge using responsibilist virtue theoretic concepts). Baehr rejects "strong conservative" views of virtue epistemology. These are views like Zagzebski's, which claim that appeal to ICVs can provide answers to traditional epistemological problems. For instance, Zagzebski provides an analysis of knowledge in terms of character trait virtues (1996). Baehr provides a strong argument against such views in (Baehr 2011).

person is responsible for, an ICV must be one acquired over time through actions of the agent. This is to be distinguished from mere cognitive faculties, skills or even talents, which are not acquired, the agent is not responsible for, and thus do not contribute to the personal worth of the subject. Genuine ICVs involve appropriate motivation: they require the subject to have a love of epistemic value (truth, knowledge, understanding, etc.). Any particular virtue is an excellence of character that allows a subject to gain an appropriate connection with the world, due to the subject's love of epistemic value. Paradigm examples of such virtues are open-mindedness and intellectual courage.

Responsibilists argue that virtue reliabilist views are mistaken in failing to appreciate the importance of character virtues to epistemology. Zagzebski (1996) and Battaly (2008), for instance, suggest that the reliabilist focus on faculties and processes makes it difficult for them to account for the way virtues such as "open-mindedness and intellectual courage impact 'high-level knowledge.' " Similarly, Roberts and Wood suggest that faculties can explain only the warrant of "beliefs on the lower end of the knowledge spectrum..."(2007, 109). "High-level" knowledge is supposed to be the kind of knowledge that is distinctively human and more difficult to obtain. This would be, for instance, knowledge gained through science, literature, and deep reflection.

Baehr (2011) offers an instructive version of this kind of criticism. He argues that reliabilists need to alter their theories in order to account for the distinctive way that character virtues can contribute to knowledge. His argument for this conclusion essentially involves two steps. First, he argues that the standard definition of a virtue employed by virtue reliabilists fails to rule out character virtues. He attributes to Greco the view that a virtue is defined as a personal trait that "plays a *critical* or *salient* role in getting the person to the truth . . . it *best explains* why a person reaches the truth" (Baehr 2011, 52). Baehr then cites a variety of cases in which he thinks various paradigm character virtues play this explanatory role in knowledge creation: a biologist who gains knowledge *because* of the two ICVs patience and focus; a reporter who learns the truth *because* of the ICV intellectual courage; and a historian who (appropriately) admits error because of intellectual honesty and humility. In each of these

cases, Baehr thinks, the character virtues play the salient, explanatory role, and should count as virtues in Greco's sense.

Roberts and Wood (2007) and Battaly (2008) make similar arguments, suggesting that character virtues are necessary to explain knowledge. Roberts and Wood, for instance, cite the example of Jane Goodall, suggesting that she could not have gotten the knowledge she did without her many character virtues: "... certain traits of character were necessary for the successful pursuit of Goodall's intellectual practices" (2007, 147).

Baehr's second step in the argument against reliabilism is to suggest that the epistemological task of judging the reliability of such character virtues is fundamentally different than judging the reliability of simple or mechanistic cognitive faculties. Character virtues have, for instance, very different conditions or environments in which they are properly employed. Faculties are only reliable in certain "friendly" environments (e.g., human vision is only reliable under certain lighting conditions). Character virtues, Baehr suggests, are most often employed in just those environments hostile to the reliability of faculties: when the situation is friendly and simple perception is reliable, a subject does not need to be intellectually determined or courageous. One's intellectual courage will be manifested in difficult situations. Character virtues, then, will be less reliable (obtain the truth less frequently), even when they are appropriately used to obtain knowledge. For this reason among others, Baehr suggests, the relevant criteria for evaluating the reliability of character virtues are quite different than the criteria for evaluating simple faculties. Thus, Baehr concludes, reliabilists must change their theories in order to account for the ways in which character virtues are reliable.

There are several reasons to take issue with this kind of challenge to virtue reliabilism. For one thing, traditional virtue reliabilist accounts do not explicitly exclude character virtues from those which can be evaluated for reliability. Sosa's treatment of the competences required for reflective knowledge in his later work is explicitly concerned with competences that are not merely innate cognitive faculties Sosa (2007, 2015).

Moreover, Baehr seems to misinterpret Greco's salience requirement when he suggests that it is a condition on what counts as a virtue.[6] Greco's requirement is that the virtue should be the salient explanation for the fact that the belief counts as knowledge. This condition is meant to constrain when a virtuous performance counts as knowledge, and was designed to help block some Gettier cases. Being salient to the explanation is not part of the definition of what a virtue is. Instead, it helps us determine when a particular belief counts as knowledge. [7]

A deeper problem facing this kind of responsibilist criticism involves the traditional way of carving up the terrain of the debate. As I note above, this traditional carving draws a distinction between "character virtues" and mere "cognitive faculties" (see Axtell 1997 and Battaly 2008). This way of understanding the terrain fails to recognize a number of important distinctions between types of competences and virtues. Moreover, I think that the notion of a "faculty virtue" is particularly unhelpful.

I will proceed to outline the distinctions that I think we should be making instead, and show how these distinctions *a*) defuse the challenges presented by Baehr and the other responsibilists, *b*) show how the reliabilist project is more fundamental than the responsibilist one, and *c*) secure the distinctively *epistemic* importance of the responsibilist project.

5.3 Three Distinctions

I am going to treat the terms "virtue" and "competence" synonymously, because I think these terms both pick out the appropriate target of virtue epistemology. I will use "ICV" to pick out the responsibilists' favored notion of virtue.[8]

I will also assume a simple definition of competence. I will make this assumption

---

[6] I think there is a similar issue with Roberts and Wood's (2007) discussion of this, and Battaly's (2008).

[7] Thanks to Megan Feeney for pointing this out to me, and to Lisa Miracchi for helpful discussion.

[8] Although my discussion proceeds in terms of competences, following Sosa, the distinctions below should be applicable to a variety of reliabilist views, especially to any version of virtue reliabilism (e.g., J. Greco 2010) or classic process reliabilism (e.g., Goldman 1979, 1998).

primarily for ease of exposition, but I take this to be a plausible starting point for a definition of competence.[9]

Competence: A competence is a disposition to succeed reliably enough at some type of performance. Each competence will thus be associated with four things:

1. A kind of performance.
2. A particular success condition.
3. A threshold for the degree of reliability required to be "reliable enough."
4. A set of environmental conditions under which reliability is judged.

Specifying these four features is necessary in order to individuate a particular competence, as well as to evaluate it.[10]

Here are the three distinctions I want to draw among different kinds of virtues or competences:

1. Constitutive vs. auxiliary competences
2. Discovery vs. justificatory competences
3. Collective (or aggregate) vs. singular (or narrow) competences.

These distinctions are meant to replace the faculty/character distinction. They offer a better way to understand the differences between the reliabilist and responsibilist projects. I will consider each in turn.

5.3.1 Constitutive/Auxiliary Distinction

Consider a subject, Clara, looking for her cat. She moves from room to room in her house, checking the various places that the cat is likely to be. She finally walks into her

[9]Cf. Sosa (2007). This definition leaves out some of the complexities in some reliabilist accounts, including those meant to help deal with Gettier problems, and with defeaters. This is done for both ease of presentation and to make the account more ecumenical. I think that the distinctions, and the account I give below of the collective auxiliary competences that are necessary for responsibilist character virtues, are consistent with a variety of views, both reliabilist and otherwise.

[10]Note that I am not claiming that such a specification is sufficient to individuate or evaluate competences.

office and sees the cat sitting on her computer keyboard. When she sees the cat, she comes to know that the cat is on the computer. There are two kinds of competences that Clara exhibits: first, a competence to find likely cat resting places, and second, a competence to visually recognize cats under standard lighting conditions. Both competences are relevant to an explanation of how Clara came to know the cat was on the computer, but in distinct ways. The first competence is an *auxiliary* competence; the second is a *constitutive* competence.

Virtue reliabilists have traditionally been concerned with *constitutive* competences, which I will define thus:

Constitutive Competence: A competence is constitutive just when its exercise is part of what constitutes a particular instance of knowledge. The successful manifestation of a constitutive competence results in knowledge.

Constitutive competences are competences of belief formation. The performance is the belief formation, the success condition is true belief, and the threshold of reliability will be some high level (at least > 50%).

Virtue reliabilists appeal to the exercise of competences in giving an analysis of knowledge. This can be viewed as a version of a traditional "JTB" account of knowledge; instead of a justification requirement, however, there is a requirement of what Sosa (2007; 2011; 2015) calls "aptness." To be *apt*, a belief must result from an exercise of competence, and there must be an appropriate connection between this exercise and the truth of the belief. Any particular instance of knowledge, then, is constituted by a true belief, the exercise of the competence that formed the belief, and the fact that an appropriate relationship holds between the truth of the belief and the exercise of competence.[11]

A paradigm example of a constitutive virtue is a fine-grained visual competence, such as Clara's "competence to form beliefs about domestic felines based on visual recognition under daylight conditions." When a subject with normal vision sees a cat sitting on a computer in front of her under appropriate conditions, she forms the

[11] For more on this idea of a competence partially constituting knowledge, see Ch. 1 of Sosa (2011).

belief, and hence acquires the knowledge, that there is a cat on the computer. Her competence to form beliefs about cats via visual recognition is a constitutive competence. The competence figures in the explanation of her having knowledge in a particular way: she knows there is a cat on the computer because her exercise of visual competence partially *constitutes* that knowledge.[12]

Constitutive competences are to be distinguished from *auxiliary* competences, which I will define so:

Auxiliary Competence: A competence that assists or enables a constitutive competence, but whose exercise is not a component of an individual instance of knowledge.

Such competences will be associated with different kinds of performances, have distinct kinds of success conditions, and have different thresholds of reliability with respect to those success conditions. Often, they involve putting subjects in a position to exercise their constitutive competences, i.e., they are competences to put a subject in a position to know. The successful manifestation of such competences will not necessarily result in knowledge.

Auxiliary competences come in a number of varieties; they can be typed according to the other distinctions described below, but we can also make more fine-grained distinctions. An important kind for my analysis are competences to *deploy* constitutive competences. (As we will see below, these are a form of auxiliary justificatory competences.) The success condition for such a competence involves effectively deploying constitutive competences.[13]

Consider again the example above, in which Clara forms the belief that a cat is on her computer. Clara has the constitutive competence to form visually-based beliefs about cats, but she also has a competence to find the right places to look. That is, she has a competence to reliably and appropriately *deploy* her constitutive competence to

[12]Constitutive competences need not be perceptual, or non-inferential in nature, however. A subject might (hopefully!) have competences for evaluating evidence before coming to a conclusion, or competences to perform logical or mathematical deduction.

[13]This means that the subject is competent both in getting into the appropriate position to deploy her constitutive competences, and is sensitive to the fact that she is in the proper position.

recognize cats. Her competence at finding likely cat resting places is a competence that puts her in a position to know where the cat is. The auxiliary competence here is a competence to use other competences; the auxiliary competence is manifested by a further exercise of (constitutive) competence. The success condition for this competence is that it deploys this cat-recognizing constitutive competence in the right locations.

In the Clara example, there is some sense in which the competence to find cat resting places is part of the explanation for how the subject comes to know the cat is on the computer. This competence is part of the explanation in a quite distinct, non-constitutive way, however: *it explains how the subject was in a position to know*.

Competences which deploy constitutive competences are just one kind of auxiliary competence. There are many others. Some are enabling competences like alertness or wakefulness, which ensure that the subject is in the right shape for possessing constitutive competences. Others are hypothesis-generating competences, or competences to ask good questions. There are also auxiliary competences which help to develop new competences over time (which we might recognize in a person, and say she is a "fast learner" or a "quick study"). Another type of auxiliary competence would be one to recognize when there are inimical circumstances or other types of defeaters present, and to stop the use of constitutive competences (these are like the reverse of deployment competences). These examples are meant to be suggestive, not exhaustive.

Whether a competence is auxiliary or constitutive may sometimes depend on the content of the belief formed. That is, a particular virtue or competence may be both constitutive of one piece of knowledge and auxiliary with respect to a different bit of knowledge. In our cat-detecting example, the competence to find cat resting places may be constitutive of knowledge of likely cat locations, while remaining auxiliary with respect to the knowledge that "the cat is now on the computer." An auxiliary competence is thus auxiliary with respect to some particular belief, by assisting the exercise of some particular constitutive competence. The exercise of the auxiliary competence must be, in some sense, prior to the exercise of the constitutive competence. This notion of priority, however, need not be temporal. There will be cases

where two competences are exercised synchronically, but the exercise of one of them is a necessary enabling condition for the other. In such a case, the former competence will be auxiliary to the latter.

This distinct way that a competence can be part of the explanation of some piece of knowledge corresponds to one of the senses of "explain" which Baehr appeals to in suggesting that ICVs can serve as the salient explanation of an instance of knowledge (2011). The same can be said for the arguments of Roberts and Wood (2007) and Battaly (2008) that knowledge cannot be gained without certain character virtues. This explains why these claims about subjects believing things "because" of an ICV are felicitous. Nonetheless, this does not mean that we cannot distinguish the ICVs from the kinds of competences that reliabilists are concerned with. The ICVs in the responsibilist's examples are serving an auxiliary role, rather than a constitutive one, in these explanations.

5.3.2 Discovery/Justificatory Distinction

The second main distinction I want to highlight is between justificatory competences and competences relevant for discovery.[14] Consider Amy, who is an experimental physicist. She comes to believe some fact about quantum fields because she came up with a hypothesis, tested it, and then (separately) judged that the test evidence was adequate to justify belief. Rory is a physics journal referee who reads Amy's work. He comes to have the same belief about quantum fields using the same justificatory procedure, but with an *entirely different method of discovery*. Amy created the hypothesis that her work confirms, and collected the relevant data; Rory merely comes to the idea and the justification from reading. In this example, Amy can be credited as having employed virtues both with respect to discovery and to justification. Rory, however,

[14]This distinction is inspired by the old distinction in the history and philosophy of science between the context of discovery and the context of justification. However, I don't mean to take on any commitments from the old debate about this distinction in the HOPOS literature. Specifically, I don't want to take on any of the baggage of the debate dealing with actual history of science vs. our current justification for a theory. Why a theory was historically accepted, for instance, isn't relevant here. The inspiration is the only connection here.

can only be credited with manifesting competence with respect to the context of justification, i.e., appropriate belief formation.[15]

Amy exhibited a competence in the "context of discovery." Her virtues in this case are relevant to inquiry in a different manner than those that, for instance, deploy constitutive competences. I will call these:

Discovery Competences: Auxiliary competences dealing with creativity and inquiry, the success conditions of which involve effective creativity, e.g., novel ideas, new experimental design, or new data.

Competences to creatively come up with new ideas and hypotheses, as well as competences to design experiments and collect data, are discovery-relevant competences. They are employed in the pursuit of knowledge, although they are not constitutive of it, nor do they deploy constitutive competences.[16] The threshold of reliability associated with such competences may be much lower than justificatory competences. For example, the degree of reliability necessary for a disposition to successfully create new hypotheses to count as a competence may be very much lower than 50%.

Such auxiliary discovery competences will account for the way responsibilists such as Baehr, Battaly, and Roberts and Wood say that agents obtain certain knowledge "because" of character virtues. Creatively coming up with new hypotheses helps explain how the subject is in a position to know. That these are a separate kind of competence is evidenced by the fact that it is easy to imagine a subject with excellent constitutive competences for evaluating evidence and arguments for a position, but who is not creative in coming up with new hypotheses which are candidates for becoming beliefs

---

[15]Compare this example with Roberts and Wood's appeal to Jane Goodall's example. They say that certain traits of character were necessary for her knowledge (2007, 109). This seems correct, but I want to suggest that the way the traits in question, like perseverance and courage, were necessary was different than the way her evidence-evaluation abilities were necessary. Her character virtues were needed to put her in a position to know. They involved auxiliary, discovery competences that enabled the formation of her knowledge.

[16]It may also be true that some of these competences can be called "constitutive," in that they may be constitutive of some successful creative process, such as in Levi's notion of abduction (1980). Thus, there is a relevant auxiliary/constitutive distinction with respect to the context of discovery. However, this distinction won't concern us here, as it is not appropriately relevant to knowledge and belief formation. With respect to the reliabilist's concerns, all discovery competences will be auxiliary. The constitutive competences that are relevant are those constitutive of knowledge.

and knowledge.

Justificatory competences, conversely, are those which take place in the "context of justification," i.e., when a subject is determining which hypothesis to believe, after all her evidence has been collected and all her hypotheses generated. In the example above, both Amy and Rory exercised justificatory competences. Such competences are directly related to successful belief formation, i.e., knowledge.[17] This notion of "directly related" is most clearly understood in contrast to the indirect way in which discovery competences assist knowledge: via discovery of new information, the creation of new hypotheses and theories, or the imagining of new ideas.

Justificatory Competences: Competences operative when the subject is forming a belief. Such competences are those that constitute knowledge, deploy constitutive competences, or otherwise directly enable knowledge.

The best way to make the distinction clear is to point to additional examples of each kind of competence.

The constitutive competences are those whose manifestations are beliefs with a sufficient degree of justification to count as knowledge. *They are thus all justificatory competences*. The virtue reliabilist project can therefore be described as elucidating the appropriate norms for constitutive justificatory competences.[18] Thus, *pace* Zagzebski, Baehr, Battaly, and the other responsibilists, reliabilists are concerned not with cognitive faculties *per se*, but with constitutive virtues. On my account, we should therefore replace a focus on so called "faculty virtues" with consideration of constitutive justificatory virtues. I will return to this point below.

In sum, all of the constitutive competences are justificatory, but not all justificatory competences are constitutive. Constitutive competences constitute knowledge by way of contributing to the justification or warrant of a belief in a particular way. In what follows, I will often refer to these as simply "constitutive competences," since all

[17] Knowledge requires that a belief be true and justified or warranted; hence the title "justificatory."

[18] Notice, however, that this category of constitutive, justificatory competences is not exhausted by the so-called "faculty virtues" that responsibilists like Zagzebski (1996) and Baehr (2011) point to as the supposed focus of reliabilists. I give an example below involving the visual competences of a botanist.

constitutive competences relevant to knowledge are justificatory.

Some auxiliary competences are also justificatory: they are exercised, as in Clara's cat example above, in direct support of constitutive competences. Competences that enable knowledge formation, or put the subject in a position to know, will count as justificatory. Auxiliary competences which deploy constitutive competences are one such variety: they are manifested in situations where the subject is trying to reliably form true beliefs and thus gain knowledge. There are also other kinds of auxiliary justificatory competences. For instance, a competence to be alert or awake may be necessary for the deployment of constitutive competences, but this alertness competence is not itself a deployment competence. A competence for alertness is merely an enabling condition for the operation of a constitutive competence. Still, a wakefulness competence is an auxiliary competence relevant to justification.[19]

The concept of a justificatory competence is not meant to account for every meaning of "justification" that is extant in the philosophical literature.[20] There are, however, at least three types of justification that are well accounted for in terms of justificatory competences. First, some subject might be highly justified in the sense that she has a greater variety of ways to come to know something, and so is less likely to miss it. This sense of justification is accounted for by appealing to the number of justificatory competences (auxiliary and constitutive) available to help her form the belief. Another sense of justification is the strength of the justification a subject has. This corresponds to the degree of reliability that the subject's constitutive competence has, i.e., in just how competent her belief-forming performance is.[21] Finally, the subject might have a justificatory competence to recognize defeaters, and to avoid utilizing a constitutive

---

[19]Consider again our example of Clara attempting to form beliefs about her cat's location. She has a constitutive competence to form visually-based beliefs about cats. She also has an auxiliary deployment competence, which reliably deploys the constitutive competence in appropriate potential locations. Furthermore, she also has a competence for remaining alert, so that both of her other competences are enabled to properly function. Thus, we can describe her as exercising two justificatory auxiliary competences in the service of her constitutive competence.

[20]Let alone in vernacular English.

[21]For these reasons, I have chosen the label "justificatory" for these competences. Note that both of these senses of "justification" refer to kinds of doxastic justification; neither corresponds to propositional justification.

competence when that competence is not in the proper environmental conditions to be reliable. This would make the subject more justified in the sense akin to safety: the subject could not easily have been wrong. Obviously, more would need to be said to support the claim that justificatory competences fully explain these intuitive notions of justification; however, I think that what has been said so far is suggestive, and is sufficient to make reasonable my choice of terminology.[22]

5.3.3 Collective/Singular Distinction

Finally, I will draw a distinction between narrow or singular competences, and competences which are composed of sets or collections of other competences. The following case will help to motivate this distinction. Mickey is a detective, father, and chess enthusiast. He is a perseverant person across a wide range of contexts and circumstances. He keeps trying, even after multiple failures, in a wide variety of intellectual pursuits. He never gives up at trying to figure out how he could have won in a chess match, even after the game is over. He keeps trying to solve crimes no matter how many times he fails to catch the criminals in question, and he spends long hours helping his children with their homework. Being good at each of these disparate activities requires a different skill set (different competences). However, there is something similar in Mickey's epistemically laudable behavior across these circumstances. I want to suggest that Mickey has a collective competence: a set of competences that operate in quite different ways in different circumstances, but with a family resemblance.

Here is how I would like to characterize the two kinds of competences to be distinguished in this section:

Singular Competences: A competence with a single one of each of the four features that are necessary for individuating competences: a single kind of performance,

[22] It is also worth noting that there might be another sense of being "justified" that corresponds to having excellent discovery-relevant competences. An investigator who is excellent at coming up with hypotheses might be more justified in coming to believe one such hypothesis than another investigator who is less likely to think of all the relevant hypotheses. The first investigator is more justified because she is less likely to miss things. This does not present a problem for my distinctions, though; as I said, justificatory competences are not meant to explain all the senses of "justification" in philosophy or ordinary language.

a single set of success conditions, a single threshold of reliability, and a single set of proper environmental conditions.

Collective Competences: A "competence" that is actually a set that is comprised of other competences. It is a family of related competences, each of which has its own four relevant features. A subject's possession of a collective competence will require her possession of large enough subset of these competences.[23]

Constitutive competences are narrow, involving a singular competence to reliably form true beliefs with respect to some subject matter.[24] Auxiliary competences, whether justificatory or discovery-relevant, may be either singular or collective (or, to restate it roughly, narrowly or broadly employed). A singular auxiliary competence is exemplified by the one in the cat example, a competence to *deploy* a small set of constitutive competences.[25]

Collective competences are *sets* of widely applicable *auxiliary* competences that are related.[26] The downstream constitutive competences are not members of these sets, but will be assisted (deployed, enabled) directly or indirectly by the auxiliary competences which are members of the set.[27] I think that the kind of broad, global

---

[23] One might begin having worries about the generality problem (See Goldman 1979, Feldman 1985, and Conee and Feldman 1998) here. As I suggest below, I think that my way of distinguishing competences may well help with the generality problem. However, the problem is one that arises for any view of knowledge that requires well-foundedness or doxastic justification (Comesaña 2006). I think there are solutions to the problem, but arguing for them is beyond the scope of this paper.

[24] There are, I take it, deep metaphysical waters here with regard to the individuation of dispositions. Furthermore, it is almost certain that any singular, constitutive competence will be describable (even reducible) in terms of the dispositions of sub-personal cognitive mechanisms. Examples of such attempted descriptions abound in vision science, for example. What is important here, however, is that the dispositions that are relevant to epistemological evaluation are singular. It might be that any visual competence can be further reduced to talk of sub-personal cognitive mechanisms. In that sense, it may be that there are a wide variety of such mechanisms, the possession of which are necessary for a subject to possess the visual competence. However, the best description for the purposes of epistemology, the person-level description, involves the subject's singular competence to successfully form beliefs (of a certain type, under certain conditions, etc.). Thus, the sub-personal does not concern us here, and I will set this point aside.

[25] Such a competence is still singular, even while deploying a set of competences, because it has just one of each of the four features: one kind of performance, one success condition, one reliability threshold, and one environmental standard.

[26] Or at least often grouped together in common vernacular, or when investigating character virtues.

[27] Since collective competences are sets, this has the result that, in some sense, the collective competence is not itself causally efficacious. Instead, the member competences are the ones which will feature in causal explanations of the subject's behavior.

character trait competences appealed to by responsibilists and virtue ethicists will involve, and even require, that the agent have families of related competences.[28] This is because of the variety of environmental conditions, success conditions, and types of performances in which the same character virtue is implicated. This explains how these virtues are "widely applicable": the collective competences they involve are exercised in a wide variety of situations. More carefully: a collective competence's *member* competences are active in a wide variety of cases. So, although competences (virtues) are not often thought to be sets, I want to suggest that the responsibilist character virtues must involve sets of competences, i.e., collective competences. This helps to explain how character virtues are "global" character traits: they require the subject's possession of a set of competences, each of which may be exercised in different circumstances, so that the set or family is implicated in many quite different kinds of activities, situations, or environments. Thus, possession of a character virtue requires possession of one or more collective auxiliary competences.[29]

Baehr's account of open-mindedness, Roberts and Wood's account of courage, and King's account of perseverance are all open to this kind of explanation. Each one of these accounts is compatible with, and well supplemented by, this notion of collective competences. Consider Baehr's notion of open-mindedness. Open-mindedness is supposed to be characterized by a willingness to transcend some cognitive standpoint (Baehr 2011, 152). Baehr gives three quite different paradigm examples of open-minded behavior in order to draw out what relates them: 1) willingness to transcend one's own beliefs, 2) ability to think openly or "outside the box," and 3) fairness in adjudicating between two opposing positions. Each of these examples illustrates what appear to be quite different competences, for the following reasons. First, these appear to be very different kinds of performances, with different success conditions, degree

[28] I think that Christine Swanton's (2001; 2003) virtue ethical view is a good example of a theory that makes this kind of thinking explicit.

[29] I suspect that we might be able to reduce character virtue talk entirely, in favor of collective auxiliary competences. That is, all there is to having an intellectual character virtue is having a certain collective competence. However, arguing for this further, more radical conclusion is beyond the scope of this paper. At the moment, I am simply arguing that character virtues involve collective competences. The benefit of this move will become clear below.

of reliability, and relevant environmental conditions. Second, it is very plausible that any subject could have a subset of the competences without having the others. Thus, open-mindedness requires that the subject possess a collective competence, consisting in a set of at least three kinds of auxiliary competence that are interestingly related.[30]

The classification of competences or virtues as being collective or singular is meant to capture something importantly distinct between the reliabilist and responsibilist projects as they have been traditionally understood. Responsibilists, much like virtue ethicists, are interested in broad, global character traits, rather than more localized or specific excellences. Traditionally, responsibilists have expressed this focus by drawing a distinction between character virtues and faculty virtues, but that distinction remains problematic for a number of reasons. When combined with the other two distinctions drawn above, the collective/singular distinction can shed more light on where the two projects differ, and why they should be complementary rather than in conflict.

The notion of a collective competence is not meant to replace the notion of a character virtue; instead, I am merely arguing that recognizing collective competences is necessary to understand ICVs, and to account for their distinctively epistemic value and purpose.[31] Nor am I arguing that we should eliminate the distinction between reliabilism and responsibilism. I only want to replace the faculty/character distinction as the primary way of accounting for the differences in the reliabilist and responsibilist projects.

---

[30] This relation could be one of mere family resemblance, or it could be something more robust, such as a genus-species relationship (i.e., "open-mindedness" could be a name for a genus consisting in several species of more narrow auxiliary competences). Either of these options is compatible with the distinction I am drawing: collective competences may come in several varieties. Thanks to Georgi Gardiner for pointing out the need to address this point.

[31] Although, as I note above, I hope to make the argument that we can reduce the notion of a character virtue to collective auxiliary competences; but that argument is beyond the scope of this paper.

5.3.4 In Support of the Distinctions

Presumably, Baehr and other responsibilists might demur about the importance of these distinctions. It is thus worth pausing to give additional consideration to the justifications for making them. First, I think the distinctions have broad appeal based simply on intuitive plausibility. It is highly intuitive that we can distinguish, on the one hand, those competences which serve as part of what provides warrant to true beliefs from, on the other hand, those other competences that simply put us in a position to know (or otherwise enable knowing). After all, everyone recognizes the distinction between knowing and being in a position to know.

The best argument for the distinctions, however, is one of explanatory power. An epistemological theory that posits the distinction between constitutive and auxiliary competences is able to make better predictions of similarity and difference between cases, and *mutatis mutandis* for the other distinctions. One example of this is in the case I cited above dealing with the experimental physicist and her journal referee. Drawing the discovery/justificatory distinction allows us to better explain the similarities and differences between the epistemic conduct of the physicist and her referee. Another example of this can be illustrated by the following three cases.

Careful Engineer: Rose is an open-minded and highly competent engineer considering designs for a bridge over a particular river. She carefully considers all the designs, keeping her mind open until she has considered each available design. She chooses design 12, competently coming to the belief that it has all of the right characteristics to make a safe bridge over this river.

Quick Engineer: Martha is another highly competent engineer working in the same office as Rose, tasked with considering the same designs. She walks into the office late, quickly looks at design 12, and competently forms the belief that it has all of the right characteristics to make it a safe bridge over this river.

Careful Intern: Donna is an intern at Rose and Martha's office, considering the same bridge designs for the river. She carefully considers all the designs, keeping her

> mind open until she has considered each available design. She chooses design 13, coming to the belief that it has all of the right characteristics to make a safe bridge over this river. However, Donna is inexperienced, and is mistaken about the safety of design 13.

A theory that recognizes the three distinctions is better able to account for the differences between these three cases. In the first two cases, Rose and Martha both come to the same piece of knowledge, that design 12 is a safe design. They both come to this knowledge by use of a constitutive competence to form true beliefs about bridge design. Rose also manifests an auxiliary competence to decide when she has considered enough designs, and to carefully weigh each design. This auxiliary competence is part of the set that partially comprises the collective competence of open-mindedness. Martha lacks (or at least fails to manifest) this auxiliary competence. Conversely, Donna has the auxiliary competence, but lacks the necessary constitutive competence to form the appropriate true belief.[32]

A theory which did not recognize the distinctions would fail to be as predictive and explanatory. For instance, a theory which focused on ICVs would provide no explanation for the similarity between the first two cases, while at the same time also providing no explanation for the difference between the second and third case. Meanwhile, a theory which focused only on constitutive competences would not explain the similarity between the first and third case.

Thus, we are well-justified in making these distinctions based both on intuitive plausibility and, more importantly, on explanatory payoff.

One might worry at this point that the definitions I have offered for the distinctions are inadequate either as necessary and sufficient conditions, or in allowing us to grasp the distinctions. One might even worry that, in particular, the constitutive/auxiliary distinction relies tacitly on a prior grasp of the faculty/character distinction that I am

---

[32]If the reader is concerned that it is the falsity of Donna's belief doing the work to distinguish the *Careful Intern* case, we could substitute a version of the case where Donna also chooses design 12, but does so only by luck; she is actually quite unreliable at choosing safe bridge designs. Thanks to Logan Douglass for helpful comments on this point.

attempting to replace.[33] I will address these concerns in turn.

First, I think that the search for non-circular necessary and sufficient conditions as a conceptual analysis is not really the appropriate methodology here, for several reasons. One reason is that I suspect that the traditional view of concepts as definitions is incorrect, and another reason is that we ought to be engaged in understanding the world itself, and not just our concepts.[34] However, even for those who are on board with the conceptual analysis project, I do not think it is necessary for my project to give the full analysis of each type of virtue. All that is necessary for my purposes is for us to be able to grasp these distinctions, and successfully apply them. To that end, I have offered definitions as characterizations, while also pointing to particular cases that I think should allow the reader to intuitively grasp the distinctions in question.

The examples provided, along with the characterizing (if not fully adequate) definition of constitutive competences, should be enough for the reader to recognize the different kinds of virtues I am suggesting exist. We should be able to recognize the different kinds of relationships it is possible for competences to have to a particular belief by considering cases like those of the cat searcher, or the engineers in this section.

Second, I do not think we must rely on any tacit appeal to the notion of faculties in order to make this distinction. For one thing, I suspect that someone unfamiliar with the traditional distinction will be able to grasp my new distinctions by appeal to the examples given. I don't think familiarity with the notion of a cognitive faculty is necessary for understanding that there are different kinds of competences in play in the engineer examples above.

More importantly, however, none of the competences displayed by the three engineers above are plausibly characterized as cognitive faculties. The engineers' competences to pick safe bridges are acquired competences having to do with recognizing and understanding a number of features of bridges. These features don't seem to be ones our cognitive powers were evolutionarily "designed" to be sensitive to. Yet it is

[33]Thanks to an anonymous referee at this journal for pressing these points.

[34]In support of these claims, see Camp (2015) and Sosa (2015) respectively.

clear there are significant differences in the three cases. There is one kind of relationship between the competence Rose and Martha share to the belief about the bridge. There is a different kind of relationship between this belief and the competence Rose has and Martha lacks. And it is clear that there is an explanatory benefit to positing that Rose and Martha have different competences, and different kinds of competences, without any of these being mere faculties.

5.4 Diagnosing the Responsibilist Critique

The foregoing distinctions divide up the terrain in the following manner. Virtue reliabilists have been concerned with the appropriate norms for knowledge and belief formation, and have appealed to the concept of a competence to give an account of those norms. For these purposes, reliabilists have been concerned with constitutive competences, all of which are singular and justificatory. Responsibilists have been concerned with the personal worth of the subject, what the subject is personally responsible for, and the value that the subject's epistemic character traits confer upon her. Accordingly, they have been concerned with character virtues which involve collective competences that are exercised in a wide range of performances and contexts. These collective competences have members which are auxiliary competences of both the justificatory and discovery-relevant kinds. Many of the analyses of particular character virtues that have been provided by responsibilists in recent works make appeal to abilities and skills that the subject must have (see Baehr (2011), Roberts and Wood 2007, Battaly 2008, and N. L. King 2014). I want to suggest that these appeals are well explained by the idea that character virtues involve collective auxiliary competences. As I will illustrate below, responsibilist analyses are thus improved by recognition of the distinctions.

The global character virtues analyzed in recent responsibilist works, such as open-mindedness (Baehr 2011), intellectual courage (Roberts and Wood 2007), and perseverance (King 2014), involve collective auxiliary competences. Nonetheless, some competences meet some of the responsibilist criteria for being an ICV, but will count

as constitutive competences. Specifically, there are constitutive competences that are acquired, skillful, and that one can be held responsible for having. I take it that the examples involving the engineers in section 3.4 are such constitutive competences. Thus, the distinction between character and faculty virtues fails to track the differences in the focus of the two projects of responsibilism and reliabilism. As I will demonstrate below, this way of carving the terrain blocks common responsibilist complaints against reliabilism.

With the distinctions in hand, an easy response to common responsibilist criticisms is available for the virtue reliabilist. Virtue reliabilism is concerned with constitutive and justificatory virtues.[35] Although the virtues that the responsibilist points to, e.g., open-mindedness and intellectual courage, do in some sense "explain" how a subject arrives at knowledge, they are not constitutive competences.

Jane Goodall would not have been able to come to her knowledge about chimps without her character virtues. But the way these virtues enable her knowledge is not by constitution. Her student, with many fewer character virtues, might have the same evidence-evaluating competence and thus come to know by appraisal of evidence painstakingly gathered by Goodall herself (recall the discussion in section 3.2). Character virtues enable knowledge because they involve auxiliary, collective, and sometimes discovery-relevant competences. Character virtues are not competences to form beliefs reliably, but rather involve competences to *deploy other belief-forming competences*. Thus, the reliabilist need not be concerned with fitting such virtues into her account of knowledge-level justification.

Consider Baehr's objection from above (in section 2) as a paradigm example of a responsibilist critique of virtue reliabilism. It is certainly correct that some ICVs may qualify as the same kind as the reliabilist appeals to, but this is because they are in fact constitutive and justificatory. This can be illustrated by consideration of a constitutive competence which meets the criteria that Zagzebski (1996) and Baehr (2011) point to

[35] I think they will be narrow or singular (in order to combat the generality problem).

as distinguishing ICVs from faculties. These criteria include that the virtue be something that is acquired and that is creditable to the subject, and which contributes to her personal worth. Such criteria are clearly met by, for instance, a botanist's competence to (non-inferentially) visually recognize particular species of plants. This is a constitutive competence, but one that is both acquired and creditable to the subject. This, however, is not the kind of "character virtue" that would require a different account of reliability. The botanist's competence is a singular competence to reliably form beliefs based on visual evidence. It has a single set of success conditions (believing truly about plant species), and requires a high degree of reliability under a certain set of environmental conditions. It is not a global trait of character that we could include in a short list of virtues necessary for the good life. This makes it much the same as other constitutive competences appealed to by the reliabilist. An account of the norms for belief-forming dispositions to count as constitutive competences requires no revision in the face of this kind of example.

In sum, some constitutive competences are acquired, skillful, and are such that we can hold subjects responsible for having them, but they are not the kinds of global character traits paradigmatically focused on by responsibilists, nor are they mere cognitive faculties. I do think they are well explained by traditional virtue reliabilist accounts, such as found in Sosa and Greco's work, so there is no need for significant revision to account for the way character virtues are deployed in inimical circumstances (see section 5.2); such virtues are auxiliary competences, which put one in a position to know, but are not the focus of reliabilism.

These responsibilist critiques of reliabilism fall short, in large part, because of a failure to recognize the distinctions above, and a reliance instead on the distinction between character and faculty virtues. This distinction fails to carve up the terrain appropriately, as it mis-characterizes the reliabilist project and its connection to responsibilism. Virtue reliabilists are not chiefly concerned with cognitive faculties as such (though the paradigm examples of faculties, like simple vision or inferential abilities, are explained well by reliabilism). Instead, reliabilism is focused on constitutive virtues, including those which are acquired and creditable to a subject's

personal worth. Moreover, alleged examples of character virtues which explain a subject's knowledge are not problematic for reliabilism because they are not constitutive competences; the way they help explain knowledge acquisition is auxiliary. Virtue reliabilists do not claim that ICVs are never reliable belief-forming competences; it is just that the responsibilists' problem examples are not constitutive competences.

Responsibilists are correct that character virtues need some other account in order to explain their relevance to reliability and knowledge. I don't think the way to do this is by appealing to different kinds of knowledge (the way that, e.g., Baehr, Battaly, and Axtell do). Instead, character virtues have a different relationship with knowledge than constitutive competences do. I argue that the way to understand this relationship is by recognizing that character virtues must at least *involve* the possession of collective auxiliary competences. In order to distinguish such competences from mere collections of dispositions, we need an account of which competences comprise the set, and of the four features associated with these member competences. However, that there is this additional need for such an account does not impugn virtue reliabilism. In fact, as I will argue below, I think that such considerations will highlight the fundamentality of the reliabilist project for epistemology.

It is also worth noting an important point of disagreement between the position I am advocating and some traditional tenets of responsibilism. In giving a more Aristotelian theory, the responsibilists are seeking a close connection between ethics and epistemology. This is illustrated by their concern with the way in which the virtues contribute to the "personal worth" of the subject, and in the way Zagzebski insists that epistemic virtues be acquired over time in a way creditable to the subject's agency. That is, responsibilists want to hold the subject responsible for her character traits, and judge the overall worth of the subject based on these traits. And this leads to a focus on a certain motivational component that they think is required: a subject should be motivated by a desire or love for truth or knowledge.[36]

This notion of personal worth and responsibility is clearly modeled on the notion

[36]Each of the responsibilists that we have discussed includes such a requirement. See the overviews of responsibilism in Axtell (1997) and Battaly (2008).

of moral worth. Indeed, I think that judgments about personal worth as appealed to by the responsibilists just are tracking the moral worth of agents. Trying to assimilate epistemic values and normativity to ethical values and normativity is a mistake, however. The two types of value/normativity can come apart. Following moral rules might lead one to acquiring less epistemic value. Conversely, one can be quite immoral but highly epistemically virtuous. I take these claims to be highly plausible, and quite defensible, although a full defense of them is beyond the scope of this paper. I will offer only a brief defense by highlighting the intuitive plausibility of this idea with some cases.

Imagine an excellent, open-minded scientist who is only in it for the money. She treats others poorly, and seeks new scientific discoveries only for personal gain. But she is open-minded and highly competent, as it turns out this is the best way for her to acquire success and the material goods that come with it.[37] Intuitively, the greedy scientist's open-mindedness does not contribute to her personal worth. Why not? Because she has it for the wrong reasons. But these are the wrong reasons, morally speaking. Epistemically speaking, she is doing as well for science as she possibly could be, and she is contributing greatly to human knowledge. Her failure here is moral.

Or, conversely, consider a person who has some strong but misleading evidence for an immoral discriminatory belief. If she ignores this evidence, she might fail to be open-minded in the epistemic sense, and thus be epistemically less than ideal, but she is behaving in a morally good way.

None of this is to say that being motivated by a desire for the truth cannot be epistemically valuable. Such a motivation might well be a component or basis for a variety of competences that make one more likely to get at the truth. Indeed, Roberts and Wood (2007) treat love of knowledge as a distinct virtue, and this coheres well with my account. My claim is simply that such motivation, and the kind of increase in personal worth that goes with it, is not necessary for distinctively epistemic value.

The responsibilists' focus on personal worth leads them to conflate the moral and

[37] See Sosa (2015) for a similar example. Thanks to Ernest Sosa for discussion on this point.

epistemic value of intellectual virtues. Being open-minded through intentional effort, for instance, makes one a better person. This seems correct, as stated; however, this notion of personal worth is ambiguous. There is a sense in which one is a better person for being open-minded, *morally speaking*, and a sense in which one is a better epistemic agent for being open-minded. It is worth keeping these two notions distinct in our thinking. For one thing, it seems plausible that one's being open-minded is intrinsically valuable from a moral perspective. Epistemically speaking, however, being open-minded is derivatively valuable: it is valuable because it leads a person to the truth and to knowledge.[38] It is this latter, epistemic sense, in which I will argue that the reliabilist project, and its target of constitutive competences, are more fundamental than responsibilism and its target collective auxiliary competences. Thus, I think we should be careful to distinguish between epistemic normativity and the ethics of belief. In what follows, I will focus on the epistemic domain, and my argument will be for the epistemic fundamentality of reliabilism. I will thus focus on the epistemic aspects of virtues like intellectual courage, leaving aside their moral value.

As I note above, the justification for the distinctions, and for the understanding of the reliabilist and responsibilist projects they allow us, is largely by way of inference to the best explanation. So, the real argument for this way of understanding the terrain is in how well it allows us to account for different kinds of examples, and for our intuitive judgments about such cases. Although I have adverted to a few cases in justifying the distinction of epistemic value from personal worth, it is largely the explanatory payoff that does the real work of justifying this way of seeing the terrain. In what follows, I will continue to elaborate on how these distinctions provide us explanatory benefits.

5.5 The Fundamentality of Reliabilism

So far, I have elucidated the three distinctions that I think are helpful for a better understanding of virtue epistemology, and then shown how the distinctions help defuse

[38] Any account that appeals to derivative value in axiology requires some solution to the "swamping problem" (L. Zagzebski 2003; L. T. Zagzebski 1996b). At least, my view here certainly does. Providing one is beyond the scope of this paper, though I am confident that some account will end up being adequate. See Pritchard et. al (2018).

certain objections to reliabilism. In this section, I will argue that these distinctions allow us to see the way the two projects of reliabilism and responsibilism complement one another.

The reliabilist project is in an important sense more *epistemically* fundamental than the responsibilist project. This can be elucidated clearly in terms of the distinctions presented above. Specifically, the virtue reliabilist is concerned with giving an account of those competences which are justificatory and constitutive of knowledge. The responsibilist project is to provide an account of intellectual character virtues. These virtues have epistemic import by way of a different relationship to knowledge: they involve possession of widely or globally active collective auxiliary competences.[39] I will argue that this project is importantly dependent on the reliabilist project.

I want to be clear from the beginning, however, that my argument is not meant to establish that the responsibilist project cannot be pursued at all without first settling all questions about constitutive competences. Rather, I want to suggest that any theory we offer about the nature of the collective, auxiliary, and/or discovery virtues must be constrained by our theory of constitutive competences (or by whatever ends up being the best theory of doxastic justification). I think it is likely that there is a great deal of fruitful work that can currently be undertaken on collective auxiliary competences. Recent analyses of virtues offered by Baehr (2011), Roberts and Wood (2007), and King (2014) are examples of this kind. This work depends, however, on what I take to be a relatively substantial amount of agreement about the nature and properties of doxastic justification.[40]

The virtue reliabilist project is epistemically more fundamental than the responsibilist project in two ways: normatively and methodologically. First, the subject matter of responsibilism, the character virtues, are *normatively* dependent on the subject

---

[39] For explanation of this notion of "widely or globally active" see section 3.4.

[40] I think there is agreement even between reliabilists and evidentialists (who have, after all, the same *explanandum* in mind). For ease of presentation, however, I will assume some form of virtue reliabilism is correct with respect to knowledge and doxastic justification. I think the argument will hold, *mutatis mutandis*, even if the appropriate account of justification turns out to be evidentialist.

matter of reliabilism, the constitutive competences. Second, responsibilism is *methodologically* dependent on reliabilism: any virtue responsibilist account will depend fundamentally on an account of (or at the very least, a sensitivity to) constitutive competences. I will address each of these dependence relations in turn.

Constitutive justificatory competences are normatively fundamental because other kinds of epistemic competences are dependent upon them for their epistemic usefulness, efficacy, and standard of evaluation. In slogan form: *Without constitutive competences, auxiliary competences would serve no epistemic purpose*. Auxiliary competences gain their distinctively expistemic value in virtue of their relation to constitutive competences. On this picture, character virtues receive their distinctively epistemic value because they involve collective auxiliary competences, which facilitate constitutive competences.

This can be illustrated by appeal to deployment competences. Deployment competences, as I have described them, clearly need competences to deploy. In order to determine whether some disposition is a competence, we need to know its success conditions. The success conditions of a deployment competence will depend, in part, on the nature of the competences they deploy. A disposition to deploy another disposition would not count as an epistemic deployment competence if the deployed disposition were not itself a competence.[41] We cannot give an appropriate account of the four features of the deployment competence without appeal to the features of the deployed competences. In particular, the success conditions of the upstream deployment competence will involve appeal to the success conditions and environmental conditions of the downstream deployed competences. The nature of the downstream competences will help determine the success and environmental conditions of the upstream deployment competence.

By similar reasoning, *discovery competences would serve no epistemic purpose without justificatory competences*. Beliefs are closely linked with behavior: if someone believes a proposition, they will behave as if that proposition is true (*ceteris paribus*). Mere

[41] Or at least the disposition deployed must be a competence most of the time the deployment competence is operative.

hypotheses do not have the same effect on behavior. If a subject were excellently creative in coming up with hypotheses and gathering evidence, but ignored the gathered evidence and formed no beliefs about these ideas, the creative excellence would serve no (epistemic) purpose. Forming a true belief is part of the success conditions of constitutive competences; this fact constrains what auxiliary discovery competences can look like.

Thus, the standards by which we evaluate auxiliary and discovery competences will depend on what the standards are for constitutive competences. This is to say, the relevant degree of reliability and environmental conditions for these auxiliary competences will in part depend on the corresponding features of the downstream constitutive competences.[42] Furthermore, collective competences are just sets of auxiliary competences, and so the fact that a collective set of dispositions is a competence will be grounded in the fact that its member auxiliary dispositions are competences, which in turn depends importantly on the fact that their downstream constitutive dispositions are competences. Thus, the constitutive competences, which are the focus of reliabilism, are fundamentally important to determining the four features of auxiliary, discovery, and collective competences. Furthermore, as I will argue at greater length below, collective auxiliary competences, of both justificatory and discovery types, must be possessed by agents as a component of character virtues. And so constitutive competences are necessary for the epistemic value of character virtues. That is, the distinctively epistemic value of character virtues is derived from their involving collective auxiliary competences (because possessing a character virtue requires possessing the collective auxiliary competence that partially constitutes it). And the value of the collective auxiliary competence is in turn derived from those constitutive competences it facilitates.

An auxiliary competence will be judged by different standards than a constitutive competence; it will have a different required degree of reliability and different

[42] Although I have framed this discussion entirely in terms of one subject's competences, there is no reason why there couldn't be a social dimension to this. It might be that one subject's competence is auxiliary to another subject's constitutive competence. At least, nothing I have said rules this out.

proper environmental conditions. Constitutive belief-forming competences must be reliable in a familiar sense: they must produce true beliefs some high percentage (at least > 50%) of the time under certain specific, favorable environmental conditions. As Baehr (2011, Ch. 4) is quick to point out, the degree of reliability required for auxiliary competences is different. However, the standards by which an auxiliary competence is evaluated are a function of how it interrelates with, and sometimes deploys, constitutive competences. An auxiliary competence, if it is one which concerns deployment of constitutive competences, will only count as a competence insofar as it successfully deploys the constitutive competences with a certain success rate (where a successful deployment is when the constitutive competence is manifested and the subject comes to know); otherwise it is a mere disposition to engage in a certain narrowly described behavior, if it is anything at all. Similarly, a discovery-centric auxiliary competence will only count as being a competence if it produces new ideas that are fit to believe, or if it provides helpful new evidence for use by justificatory competences.

Intellectual character virtues involve, at least as components, collective auxiliary competences. They are partially composed of more locally applicable auxiliary competences, and possession of auxiliary competences depends on possession of constitutive competences. Thus, the normative question of whether a set of dispositions really is a collective *competence* depends on whether the members of the set are competences. And whether an auxiliary disposition counts as a *competence* will depend on whether it assists some genuine constitutive competence.

To illustrate this point about the normative question of whether a disposition counts as a competence, consider again auxiliary deployment competences. Whether some disposition, A, *competently* deploys another disposition, β, depends on whether β is reliable or effective enough under the circumstances. Moreover, A must be sensitive to β's reliability. If β is not often enough reliable when deployed by A, then the disposition A may not be a competence or virtue at all. Of course, auxiliary competences will sometimes be competences to deploy constitutive competences in difficult circumstances. This might mean that β need not be highly reliable by itself under the relevant

circumstances. Instead, it might be that A is a competence to deploy constitutive competences β, γ, δ, and ϵ whenever certain, epistemically inimical circumstances arise. In order for A to count as a competence in this case, it must reliably deploy these competences in such a way that the end result is that at least one of β, γ, δ, or ϵ is ultimately successful in forming a justified belief. This might be accomplished by A deploying each of β, γ, δ, or ϵ in turn.

This kind of understanding would provide a nice addition to our theory of a character virtue like perseverance, as analyzed, e.g., by King (2014). Possessing perseverance involves possession of an auxiliary competence A. Whether A is a competence will depend on the degree of reliability of β, γ, δ, or ϵ, the conditions under which they are in fact reliable, how they interact with one another, and how successfully they are deployed by A.

A subject need not *know* how reliable a constitutive competence is, nor be able to make the judgments mentioned above explicitly, in order to possess an auxiliary competence related to them. However, a subject must be *sensitive* to the reliability of the relevant constitutive competences in order to have an auxiliary competence. Such a sensitivity will be part of what it is to have the auxiliary competence.

Thus, the reliabilist project is normatively fundamental to the responsibilist one. Intellectual character virtues count as epistemically valuable because of their relationship to constitutive competences. This is because character virtues involve possession of collective auxiliary competences, and evaluating such things as competences involves appeal to constitutive competences.

One might object that there is another way of seeing the normative priority here: that auxiliary competences are prior. An agent *uses* her constitutive competences for her own ends, via her auxiliary competences. I think this is probably a good way of describing an agent's actions in many cases. However, I don't think this is the right way to understand the normative relationship between the different kinds of virtue, because of the way that the constitutive and auxiliary competences can come apart. A well-functioning constitutive competence can provide knowledge in the absence of an auxiliary competence, and knowledge (or the truth it ensures) is plausibly at the root

of epistemic value. Conversely, an auxiliary competence in the absence of constitutive competences does not have such a link to something epistemically valuable. Now, auxiliary competences are valuable, I suggest, precisely because they are so important in enabling our constitutive knowledge-producing competences. But they aren't strictly speaking necessary (even in cases like those of the engineers in section 3.4).

Reliabilism is also methodologically fundamental to the responsibilist project. Giving an account of the relevant kind of performances, success conditions, degree of reliability, and proper environmental conditions for collective auxiliary competences depends on having an idea (or at least some assumptions) about the requirements of the downstream constitutive competences. For reasons similar to why possession of auxiliary competences depends on possession of constitutive competences, giving an account of the standards for constitutive competence is a necessary prerequisite for giving a complete account of any auxiliary competences. More usefully, any account that we offer now should be constrained by our current account of constitutive competences. That is, when investigating the four features of any auxiliary competence, the epistemologist (or psychologist) should be sensitive to the features of the downstream constitutive competences relevant to the auxiliary competence in question. Therefore, the virtue reliabilist project of explaining doxastic justification in terms of competence should constrain the theories we offer about auxiliary competences of all kinds, including the sorts of collective auxiliary competences that have most concerned the virtue responsibilists. This is true regardless of whether the auxiliary competences are justificatory deployment competences, such as in the cat example above, or whether they are discovery competences to propose new hypotheses.

The *manner* in which accounts of these different kinds of competence will be constrained, however, may be quite different. For instance, our understanding of constitutive competences means that auxiliary justificatory competences will require a certain kind of success rate or reliability. A discovery competence, however, need not have a high degree of reliability in order to count as a competence, at least not in the sense of needing to produce ideas which come to be beliefs a majority of the time. A hypothesis creation competence might produce false hypotheses 99.9% of the time, yet

still be an example of an incredible competence if the successful ideas are excellent, or if the generated hypotheses are very creative. Note also that collective competences may consist of both justificatory and discovery competences, as in the examples of paradigm character virtues. Thus, any collective competence should be evaluated on the basis of the individual auxiliary competences comprising its membership.

In sum, any account of auxiliary competences is dependent upon, and should be constrained by, an account of constitutive competences. In order to illustrate this, in the following section I will provide an example account of a collective auxiliary competence, highlighting the fundamental importance of the constitutive competences and the reliabilist account of them. Following Baehr (2011) and Roberts and Wood (2007), I have chosen to focus largely on an important exemplar character virtue, intellectual courage. This will highlight how appeal to the distinctions offers significant help to the responsibilist project, while not serving as any sort of competition for, or critique of, these traditional accounts.

Before moving on to the example, however, it is worth taking stock of the projects of this paper. First, I have attempted to provide a set of distinctions that more appropriately divides the terrain of virtue epistemology. I then argued that these distinctions show the flaw in standard responsibilist criticisms of reliabilism. These two goals comprise the first, weaker project of the paper. In this current section, I have argued that there is a certain sense in which the reliabilist project of elucidating (and appealing to) constitutive competences is more fundamental than the responsibilist project. I have suggested a way of understanding the character virtues that concern responsibilists by appeal to the idea of collective auxiliary competences. This project is separable from the foregoing ones, though I think it follows quite nicely from them.

5.6 Intellectual Courage

In support of my claim about the fundamentality of reliabilism, in this section I will apply the distinctions to the paradigmatic character virtue of intellectual courage. In order to increase the persuasive power of the argument, I will use Baehr's account of

the virtue as a starting point. I will show that his theory is helpfully clarified by application of the distinctions, and that in fact Baehr's account is significantly improved when we appeal to these distinctions. Then, I will argue that Baehr's project is appropriately constrained by and ultimately dependent on answers to the virtue reliabilist project. I focus on Baehr's account for ease of exposition, and because I take it to be an excellent recent example of a responsibilist analysis of a character virtue. The distinctions, and the connection to constitutive competences that I argue for here, are also applicable to a variety of other recent responsibilist analyses of virtues, for instance those by Zagzebski (1996), Roberts and Wood (2007) and King (2014).[43]

The character virtue intellectual courage (IC) requires possession of a collective auxiliary competence. That is, being intellectually courageous requires having a collection of auxiliary competences that share some important, characteristic traits or properties.[44] Baehr (2011) defines IC as follows:

> **(IC)** Intellectual courage is the disposition to persist in or with a state or course of action aimed at an epistemically good end despite the fact that doing so involves an apparent threat to one's own well being. (2011, 177).

This definition can be broken down into two essential parts, what Baehr calls the *context* and *substance* of IC (2011, 169). The context aspect of the definition picks out the relevant circumstances when IC can be manifested, i.e., those circumstances when there is a threat to the subject's well-being conditional on pursuing some good. The substance aspect of the definition picks out the kinds of actions a subject can engage in or pursue courageously. Baehr spends a significant amount of time elucidating the context aspect of IC. He arrives at the notion of "apparent threat to well being" after analyzing several alternative formulations. He points out that the subject in question need not have any actual fright affect associated with the danger, nor need there be

---

[43] As discussed in the last section, I think it is important to keep the aspects of courage that are morally valuable distinct from those of epistemic value. As such, I will focus on intellectual courage as an epistemic virtue, and not on cases of intellectual courage where one is (only) morally creditable for pursuing truth in the face of danger.

[44] Or, as I mentioned above, perhaps these competences are species of the same genus called IC, but I will focus on the former option.

a high likelihood of the danger manifesting. However, there must be some sense in which the subject recognizes some possibility or threat of harm. I take Baehr's account of this aspect of IC to be quite plausible, aptly representing the kind of fruitful positive work that can be accomplished on the responsibilist project.

Baehr's difficulties with the positive account of the "substance" of IC, however, illustrate the usefulness of the distinctions, and the fundamental importance of the reliabilist project. With respect to the substance of IC, Baehr suggests that "at a certain level there is no answer to this question, for the substance of intellectual courage is to a significant extent indeterminate" (2011, 173). The problem for Baehr is that the instances of intellectual courage are many and widely varied, and seem to have little in common. The only common features are the involvement of the pursuit of intellectual value and the context conditions described above. He notes that this differs from many other virtues which have some particular kind of activity necessarily associated with them. Here, Baehr's account will benefit from an application of the distinctions above.

According to the current account, IC involves a collective, auxiliary competence. It requires that the subject possess a collection of dispositions to deploy various other constitutive and auxiliary competences. What members of this collection have in common is that they manifest under the circumstances that we have been calling the proper context of IC. Here is an amended account:

> **(IC*)** Intellectual courage is a disposition which involves a *set of* auxiliary competences *to deploy* certain other epistemic competences. The members of the set are similar in that they are competences to weigh apparent threat to one's own well being, and to deploy relevant downstream competences despite significant threat.[45]

Including the term *competence* builds in the fact that the deployed dispositions will

[45]The most straightforward amended definition would be to simply identify IC with a collective auxiliary competence, effectively reducing talk of character virtues to talk of such competences. As I note in footnote 29, I think this is probably the right way to go. However, that would require additional argument in favor of such a reduction, which is a separate project for future research. So I am here only endorsing the weaker claim, that IC involves or necessarily requires the possession of a collective auxiliary competence.

be aimed, directly or indirectly, toward some positive epistemic value. According to the virtue reliabilists, the end in question will be truth (or perhaps knowledge). IC is thus understood to involve a set of competences to deploy other competences, and we will see below that this helps to better explain the variety of instances of IC. Some competences that are deployed by members of IC's collective auxiliary competence will be constitutive, while others will be auxiliary, and these may be either justificatory or discovery competences.

Thus, by appeal to the appropriate distinctions, it becomes clear that the substance of IC is not significantly indeterminate. While it may be a complicated matter to categorize each of the competences comprising IC, as well as all of the competences deployed by IC, our distinctions at least give us a map of what this kind of elucidation would look like, and takes some of the mystery out of the nature and structure of intellectual courage.[46]

Furthermore, the distinctions clarify the way in which IC is aimed at *epistemic* goods or intellectual ends: it facilitates the functioning of constitutive, reliabilist competences to get at the truth. When a member of IC is manifested, it deploys some other competence; when this competence is constitutive, it reliably arrives at the truth. When the competence deployed is auxiliary, it will also be aimed at facilitating the truth, though more or less indirectly. For instance, a member competence of IC might deploy an auxiliary discovery competence for discovering new hypotheses, which will later be used by constitutive competences as material to reliably form new beliefs about. As I have suggested, knowing when a disposition *A competently* deploys another disposition β requires knowing something of the nature of β; specifically, the success conditions of β, its required degree of reliability, and its environmental conditions. In the case of IC, knowing whether an act of continuing inquiry (for instance)

[46]Notice also that there are two distinct possible projects of elucidation that we might engage in. First, we might elucidate all of the competences that comprise IC as such, meaning all of the competences that any subject may have which would count as part of her IC. Second, we might attempt to give an account of all of the relevant competences that an actual intellectually courageous subject possesses. Presumably, we might count a subject as having IC even if she does not have every competence which could be part of that intellectual virtue, so the two projects come apart.

is courageous will depend on knowing whether the kind of discovery competence deployed in the inquiry is effective (enough) under the circumstances to warrant the risk. Otherwise, the auxiliary disposition is not one of courage but of rashness.[47]

Baehr has trouble elucidating the substance aspect of IC because he fails to recognize the fact that it requires a collective auxiliary, deployment competence. Seeing where Baehr runs into trouble helps illustrate the fundamental importance of the reliabilist project, concerned as it is with constitutive competences, in giving an account of IC. It is precisely the reliabilist account of constitutive competences, along with the distinctions outlined above, that is missing from his account of the substance of IC. In an attempt to bolster what he sees as the vague and indeterminate account of the substance of IC, Baehr picks out as examples three kinds of disparate activities in which IC can be said to operate. I think these are apt examples, and their aptness actually becomes clearer once we apply the distinctions, and adopt the virtue reliabilist account of constitutive competences. I will focus on two of these kinds of activity.

The first kind of activity Baehr picks out is the quite general category of "inquiry" (2011, 173). A subject can be seeking to find the truth (i.e., trying to obtain knowledge about a particular subject matter) even in the face of a threat of harm.[48] A subject might begin a new inquiry under threat, or sustain an inquiry when a new threat arises, or even abandon a line of inquiry when there is a threat to her well-being in doing so. The current account offers a clear way of understanding this type of IC. IC requires a collective competence, which consists in a set of competences. A subject who possesses this inquiry-relevant part of IC has a competence which is a member of the set IC, call it competence $\Gamma \in IC$. This Γ is an auxiliary competence to decide whether the subject should deploy further discovery competences, call them β and γ, in pursuit of some particular epistemic goal. Thus, Γ consists in a disposition to appropriately weigh the threat to the subject's well-being against the potential benefits

[47] For more on the importance of distinguishing rashness from courage, cf. Roberts and Wood 2007, Chapter 8.

[48] Roberts and Wood's appeal to the example of Jane Goodall is also relevant, here. She "subjected herself indiscriminately to the dangers of the forest"(2007, 224), not recklessly, but because of the value of the inquiry.

of deploying β and γ, as well as any deontic duties the subject may have for such deployment. A subject who has Γ will be appropriately sensitive to the situation, and so will deploy her β and γ often enough even when threatened.

We can illustrate this via an example. Amy, an investigative reporter, is considering whether to cover a protest happening in Egypt. Her editor, the police, and the U.S. State Department have warned her that there is a significant threat of harm if she covers the protest (from police, counter-protesters, and even perhaps professional backlash). She recognizes that there is danger. She is an astute observer, and is competent at gathering evidence with her eyes and camera (call this competence β). She is also adept at coming up with hypotheses about what she is witnessing: for instance, thinking up the idea that the counter-protesters are really government shills in disguise (call this competence γ). She is intellectually courageous, and has a particular competence to decide whether and when it is appropriate to risk danger (competence Γ). Her Γ competence allows her to reliably judge when the benefits of reporting, and her duties to do so, outweigh the significant danger she faces in deploying her competences β and γ. Thus, in this case she aptly goes ahead and attends the protest, using her keen eyes and mind to help her cover the story and fulfill her duties as a journalist, even when threatened with harm.

It is impossible to judge whether Amy possesses IC without also having some understanding of the four features of Amy's competences β and γ. Furthermore, Amy would not be intellectually courageous without herself having some sensitivity to the degree of reliability of these deployed competences, and the circumstances under which they are reliable. She must have some hope of achieving the epistemic end in question in order for her manifestation of Γ to count as competent. Otherwise, she is merely being rash, rather than courageous.[49]

---

[49]Perhaps one might be concerned here that there is no room for rashness in a purely epistemic version of intellectual courage. That is, from a *purely* epistemic viewpoint it might seem that it is always better to continue inquiry in the face of danger. It is only when we admit moral or practical considerations, the objection goes, that it seems like the epistemic benefit of further inquiry can be outweighed by the danger. I am not convinced of this, however. For one thing, if there really is *no hope* of epistemic benefit from further inquiry, then it really does seem rash to face danger for no reason. It seems like a failure to recognize a *lack* of epistemic value. Moreover, I think there are probably other cases in which the benefit is just not adequate to justify the danger. In such cases, I am tempted to suggest that a small chance

The second example of the substance aspect of IC that Baehr appeals to is belief formation or maintenance: "Any intellectually courageous person might also, it seems, adopt or maintain a belief that he regards as intellectually credible or justified despite the fact that doing so involves certain risk or potential harm" (Baehr 2011, 174). An intellectually courageous person plausibly forms beliefs according to epistemic standards even when faced with threat of harm.[50] This example of the substance of IC is well-explained by application of our distinctions. Some of the members of the set of competences which (partially) comprise IC will be competences to deploy constitutive competences, i.e., reliable belief-forming dispositions. Consider competence $\Phi \in IC$. Φ is an auxiliary justificatory competence, a competence to deploy constitutive competences χ, ψ, and a variety of others. Further, let's suppose χ and ψ are competences to form beliefs out of hypotheses based on evidence in the subject's possession. Φ is then a competence to deploy other competences like χ and ψ even under circumstances when deploying them incurs significant risk to the subject. Again, whether Φ counts as a competence will depend on the subject's sensitivity to the conditions under which, and the degree to which, χ and ψ are reliable.

The aptness of the above analysis can be illustrated by appealing again to the case of Amy. After Amy has attended the protest for long enough, and gathered appropriate evidence, she is in a position to form beliefs about the protest. However, what the evidence she has gathered strongly supports is the belief that her own government, and even the newspaper that employs her, are complicit in atrocious crimes committed against civilians. Coming to believe this would cause Amy to have to radically revise her understanding of her own life, projects, and goals. It would involve significant risk of harm to her own well-being, via her mental health and future employment. Nonetheless, in the face of this risk, she forms the appropriate belief based on the evidence that her employer and government are complicit.

---

of uncovering evidence or otherwise gaining value through inquiry could be outweighed by the danger because the danger would prevent us from gaining other knowledge later. Or one might inappropriately risk losing knowledge from death or other damage. So, I think that it can be rash and not intellectually courageous to engage in risky behavior for slight epistemic gain.

[50] I will sidestep the issue of doxastic voluntarism. I think Baehr is correct in suggesting that the virtue theoretic account of IC will survive even a pretty robust version of doxastic involuntarism.

In forming this belief, Amy manifests an auxiliary competence Φ when she deploys her constitutive competences χ and ψ. This competence Φ is (like Γ, in the previous example) a competence to weigh the risks, benefits, and duties relevant to the situation, and to deploy the relevant competences, in this case the evidence-evaluation competences χ and ψ. Thus, when Amy forms her belief about her employer and government, she does so courageously because she manifests Φ, which (in this case) deploys χ and ψ. Again, we see the fundamental importance of the constitutive competences to an understanding of the character virtue of intellectual courage, and thus we see the fundamental importance of the virtue reliabilist project to the responsibilist project.

Thus, on the present account, IC requires possession of a set of competences. A subject will possess some subset of this set, and if the subset is large enough (or if the importance of certain members is greater than others, if the subset is central enough) she can be considered intellectually courageous *tout court*. If her possessed subset is too small, she may just be intellectually courageous with respect to a few areas. Furthermore, it is intuitively plausible that one can be more or less intellectually courageous in two ways. First, a subject may be extremely courageous in one particular area (like Amy's courage in investigating). Second, a subject may be courageous across a wide range of circumstances. The present account can happily accommodate this intuition by appeal to the collective auxiliary competence necessary for IC. A subject can have a single disposition, $\Phi \in IC$, which is highly competent. Or, she may possess a large subset of the members of IC.[51]

A full account of the collective auxiliary competence associated with intellectual

[51] It is worth noting that this last feature of my account makes it compatible with the situationist literature in psychology (see Doris, Stich, Phillips, and Walmsley 2017). Psychological experiments tell us that many people's behavior can be altered by small changes to their environment, and this casts doubt on the notion of global character traits. By explaining global character virtues in terms of sets of auxiliary competences, my account can easily allow for this. What has happened in the psychology experiments is that the environmental conditions have been changed.

Furthermore, I think this might help defuse a complaint that a responsibilist might raise against my account. That is, intuitively, such character virtues are *unitary* features of a subject. However, given the aforementioned situationist psychology literature, this intuition (like many psychological intuitions) turns out to be misguided. My view can easily account for this, while the traditional "unitary" notion of character virtues cannot. Thanks to Eliabeth Fricker for helpful comments on this point.

courage would require filling out the set of competences of which it is composed, perhaps with an appropriate taxonomy, and (hopefully) with a spate of useful generalizations. This is a significant and worthwhile project, and I think that this is precisely the kind of useful and philosophically interesting account that virtue responsibilism is concerned with. What we have seen, however, is that this project requires some account of constitutive competences, and this is precisely what the virtue reliabilist is seeking to provide.

Any account of intellectual courage should thus be constrained by (at least what is common to) our best accounts of constitutive competences. Thus, the examination of IC in light of our distinctions has helped to illuminate the fundamental importance of the reliabilist project to responsibilism.

Conclusion

There has been a perception of conflict between virtue reliabilists and virtue responsibilists. If what I have argued above is correct, then this perception is misguided: there need be no such conflict among virtue epistemologists. The responsibilist project is an important and potentially fruitful area of philosophical research, but it is not attempting to explain the same things as virtue reliabilism. Once we apply the distinctions between types of virtues, it becomes clear that the two projects are after different *explananda*. Moreover, the responsibilist project importantly depends on the reliabilist project, and the latter is therefore more epistemically fundamental. Let me be explicit that this is not any form of criticism or belittlement of responsibilism: biology fundamentally depends on physics, but this is hardly a complaint against biology. All I want to argue is that the continued sense of conflict between the two camps should be swept away.

I think it is also worth pointing out that the distinctions I draw above could be useful quite apart from this in-house debate among virtue epistemologists. For instance, the distinctions may be helpful in the debate about the generality problem.[52]

[52] See Comesaña (2006), Beebe n.d., and Conee and Feldman 1998.

The distinctions can help the virtue epistemologist narrow down the number of competences or dispositions that might be the relevant one for evaluating the reliability of a particular belief formation. That is, the distinctions help to cut down on the range of generality that needs to be considered. For example, Baehr appeals to intellectual character virtues as being virtues that best explain individual cases of belief formation. This illustrates the way in which highly general virtues or dispositions can be (I think mistakenly) included in those that might be relevant for evaluation of reliability (i.e., as contributing to the generality problem). Once we apply the distinction between auxiliary and constitutive competences, however, we can rule out auxiliary competences like IC as being relevant to evaluating the reliability of a particular belief formation. Thus, the distinctions can actually help to narrow down the set of candidate dispositions for reliability evaluation.

Moreover, one need not think that virtue reliabilism is the right account of doxastic justification, or knowledge-level warrant, in order to make use of these distinctions or my account of how intellectual character virtues require certain competences. Anyone who takes the notion of epistemic virtue to be interesting or significant can make use of this account.

References

Alston, W. M. (1996, January). Belief, Acceptance, and Religious Faith. In J. Jordan & D. Howard-Snyder (Eds.), *Faith, Freedom, and Rationality: Philosophy of Religion Today* (p. 3-27). Rowman & Littlefield.

Alvarez, M. (2016). Reasons for Action: Justification, Motivation, Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 ed.).

Arpaly, N. (2017). Epistemology and the Baffled Action Theorist. , *manuscript*.

Axtell, G. (1997). Recent work on virtue epistemology. *American Philosophical Quarterly*, *34*(1), 1–26.

Baehr, J. S. (2011). *The inquiring mind: On intellectual virtues and virtue epistemology*. Oxford University Press on Demand.

Barnes, E. C. (1999). The Quantitative Problem of Old Evidence. *British Journal for the Philosophy of Science*, *50*(2), 249–264.

Battaly, H. (2007). Battaly: Intellectual virtue and knowing one's sexual orientation. In *Sex and ethics: Essays on sexuality, virtue, and the good life.* (p. 149-161). New York: Palgrave Macmillan.

Battaly, H. (2008). Virtue Epistemology. *Philosophy Compass*, *3*(4), 639-663. doi: 10.1111/j.1747-9991.2008.00146.x

Beebe, J. R. (n.d.). The Generality Problem, Statistical Relevance and the Tri-Level Hypothesis. *Noûs*, *38*(1), 177-195. doi: 10.1111/j.1468-0068.2004.00467.x

Berker, S. (2013). Epistemic teleology and the separateness of propositions. *Philosophical Review*, *122*(3), 337–393.

Bonner, B. L., Baumann, M. R., & Dalal, R. S. (2002). The effects of member expertise on group decision-making and performance. *Organizational Behavior and Human Decision Processes*, *88*(2), 719–736.

Bright, L. K. (2016). On Fraud. *Philosophical Studies, forthcoming*, 1–20.

Broome, J. (2015, October). Reason versus ought. *Philosophical Issues*, *25*(1), 80-97. doi: 10.1111/phis.12058

Camp, E. (2015). Logical Concepts and Associative Characterizations. In E. Margolis & S. Laurence (Eds.), *The Conceptual Mind* (p. 591-622). MIT Press.

Cherniak, C. (1983). Rationality and the structure of human memory. *Synthese*, *57*(2), 163–186.

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *Philosophical Review*, *116*(2), 187–217.

Christensen, D. (2009, September). Disagreement as Evidence: The Epistemology of Controversy. *Philosophy Compass*, *4*(5), 756-767. doi: 10.1111/j.1747-9991.2009.00237.x

Christensen, D. (2010a, July). Higher-Order Evidence1. *Philosophy and Phenomenological Research*, *81*(1), 185-215. doi: 10.1111/j.1933-1592.2010.00366.x

Christensen, D. (2010b). Rational Reflection. *Philosophical Perspectives*, *24*(1), 121–140.

Cohen, L. J. (1989a, July). Belief and Acceptance. *Mind*, *98*(391), 367-389.

Cohen, L. J. (1989b, January). What Use Are Beliefs That We Do Not Take to Be Warranted? *Analysis*, *49*(1), 7. doi: 10.2307/3328887

Cohen, L. J. (1995). *An Essay on Belief and Acceptance* (Reprint edition ed.). Oxford: Oxford University Press.

Comesaña, J. (2006). A Well-Founded Solution to the Generality Problem. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *129*(1), 27-47.

Conee, E., & Feldman, R. (1998, January). The Generality Problem for Reliabilism. *Philosophical Studies*, *89*(1), 1-29. doi: 10.1023/A:1004243308503

Crupi, V. (2015). Confirmation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2015 ed.).

Decker, J. (2014). Conciliation and Self-Incrimination. *Erkenntnis*, *79*(5), 1099–1134.

De Cruz, H., & De Smedt, J. (2013). The Value of Epistemic Disagreement in Scientific Practice. The Case of Homo Floresiensis. *Studies in History and Philosophy of Science Part A*, *44*(2), 169–177.

DeRose, K. (2016). *The Appearance of Ignorance: Knowledge, Skepticism and Context*, (Vol. 2). forthcoming: Oxford University Press.

Doris, J., Stich, S., Phillips, J., & Walmsley, L. (2017). Moral Psychology: Empirical Approaches. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.). Metaphysics Research Lab, Stanford University.

Earman, J. (1992). *Bayes or bust?: A critical examination of Bayesian confirmation theory*. Cambridge, Mass: MIT Press.

Easwaran, K. (2013). Expected Accuracy Supports Conditionalization—and Conglomerability and Reflection. *Philosophy of Science*, *80*(1), 119–142.

Easwaran, K. (2016). Dr. Truthlove Or: How I Learned to Stop Worrying and Love Bayesian Probabilities. *Noûs*, *50*(4), 816–853.

Eells, E. (1985). Problems of Old Evidence. *Pacific Philosophical Quarterly*, *66*(3), 283.

Eells, E., & Fitelson, B. (2000, December). Measuring Confirmation and Evidence. *The Journal of Philosophy*, *97*(12), 663. doi: 10.2307/2678462

Egan, A. (2008a). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47–63.

Egan, A. (2008b, July). Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, *140*(1), 47-63. doi: 10.1007/s11098-008-9225-1

Elga, A. (2007, September). Reflection and Disagreement. *Noûs*, *41*(3), 478-502. doi: 10.1111/j.1468-0068.2007.00656.x

Elga, A. (2010). How to disagree about how to disagree. In R. Feldman & T. Warfield (Eds.), *Disagreement*. Oxford University Press.

Elga, A., & Rayo, A. (2015). *Fragmentation and Information Access*.

Elgin, C. Z. (2010). Persistent Disagreement. In R. Feldman & T. A. Warfield (Eds.), *Disagreement*. Oxford University Press.

Faye, J. (2014). Copenhagen Interpretation of Quantum Mechanics. In E. N. Zalta (Ed.), *The*

Stanford Encyclopedia of Philosophy (Fall 2014 ed.).

Feldman, R. (1985). RELIABILITY AND JUSTIFICATION. *The Monist*, *68*(2), 159-174.

Finlay, S., & Schroeder, M. (2015). Reasons for Action: Internal vs. External. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2015 ed.).

Firth, R. (1981). "Epistemic Merit, Intrinsic and Instrumental&Quot. *Proceedings and Addresses of the American Philosophical Association*, *55*(1), 5–23.

Fitelson, B. (2016). *Coherence*. manuscript.

Fitelson, B., & Easwaran, K. (2015). Accuracy, Coherence, and Evidence. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology Volume 5*. Oxford: Oxford Studies in Epistemology.

FitzPatrick, W. J. (2004, January). Reasons, Value, and Particular Agents: Normative Relevance without Motivational Internalism. *Mind*, *113*(450), 285-318. doi: 10.1093/mind/113.450.285

Frances, B. (2010). The reflective epistemic renegade. *Philosophy and Phenomenological Research*, *81*(2), 419–463.

Frances, B. (2013). Philosophical renegades. In D. Christensen & J. Lackey (Eds.), *The Epistemology of Disagreement*. Oxford University Press Oxford.

Frances, B., & Matheson, J. (2018). Disagreement. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). Metaphysics Research Lab, Stanford University.

Frankish, K. (2004). *Mind and Supermind*. Cambridge: Cambridge University Press.

Friedman, J. (2017). Why Suspend Judging? *Noûs*, *51*(2), 302–326.

Frigg, R., & Nguyen, J. (2016). Scientific Representation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.).

Garber, D. (1983). Old Evidence and Logical Omniscience in Bayesian Confirmation Theory. In J. Earman (Ed.), *Testing Scientific Theories* (Vol. 10). Minneapolis: University of Minnesota Press.

Geil, D. M. M. (1998, July). Collaborative Reasoning: Evidence for Collective Rationality. *Thinking & Reasoning*, *4*(3), 231-248. doi: 10.1080/135467898394148

Gert, J. (2009). Desires, Reasons, and Rationality. *American Philosophical Quarterly*, *46*(4), 319–332.

Gillies, A. S. (2001). *Rational Belief Change* (Unpublished doctoral dissertation). Dissertation.

Glymour, C. (1980). *Theory and Evidence*. Princeton University Press.

Goldberg, S. (2013a). Defending Philosophy in the Face of Systematic disagreement. In D. Machuca (Ed.), *Disagreement and Skepticism*. (pp. 277–94). New York: Routledge.

Goldberg, S. (2013b, May). Inclusiveness in the face of anticipated disagreement. *Synthese*, *190*(7), 1189-1207. doi: 10.1007/s11229-012-0102-2

Goldberg, S. (2015). *Assertion*. Oxford University Press.

Goldman, A. I. (1979). What is Justified Belief? In *Justification and Knowledge* (p. 1-23). Springer, Dordrecht. doi: 10.1007/978-94-009-9493-5_1

Goldman, A. I. (1986). *Epistemology and Cognition*. Harvard University Press.

Goldman, A. I. (1998). Reliabilism.
doi: 10.4324/9780415249126-P044-1

Gordon, J. (1997). John Stuart Mill and the "Marketplace of Ideas". *Social Theory and Practice*, *23*(2), 235-249.

Greaves, H. (2013, October). Epistemic Decision Theory. *Mind*, *122*(488), 915-952. doi: 10.1093/mind/fzt090

Greaves, H., & Wallace, D. (2006). Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*, *115*(459), 607–632.

Greco, D. (2014). Iteration and Fragmentation. *Philosophy and Phenomenological Research*, *88*(1), 656–673.

Greco, J. (2010). *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge University Press.

Hartmann, S., & Fitelson, B. (2015). A New Garber-Style Solution to the Problem of Old Evidence. *Philosophy of Science*, *82*(4), 712–717.

Hawthorne, J. (2005). Degree-of-Belief and Degree-of-Support: Why Bayesians Need Both Notions. *Mind*, *114*(454), 277–320.

Howson, C. (1991). The 'Old Evidence' Problem. *British Journal for the Philosophy of Science*, *42*(4), 547–555.

Howson, C., & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing.

Huber, F. (2013a, July). Belief Revision II: Ranking Theory. *Philosophy Compass*, *8*(7), 613-621. doi: 10.1111/phc3.12047

Huber, F. (2013b, July). Belief Revision I: The AGM Theory. *Philosophy Compass*, *8*(7), 604-612. doi: 10.1111/phc3.12048

Jeffrey, R. C. (1983). Bayesianism With A Human Face. In J. Earman (Ed.), *Testing Scientific Theories* (pp. 133–156). University of Minnesota Press.

Jeffrey, R. C. (1990). *The logic of decision*. Chicago: University of Chicago Press.

Jehle, D., & Fitelson, B. (2009). What is the "equal weight view"? *Episteme*, *6*(03), 280–293.

Jenkins, C. (2007). Entitlement and Rationality. *Synthese*, *157*(1), 25–45.

Johnson, A. (2016). *Notes on Fragmentation*.

Joyce, J. M. (1998). A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, *65*(4), 575–603.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Kahneman, D. (2013). *Thinking, Fast and Slow* (1st edition ed.). New York: Farrar, Straus and Giroux.

Kaplan, M. (1981a). A Bayesian Theory of Rational Acceptance. *The Journal of Philosophy*, *78*(6), 305-330. doi: 10.2307/2026127

Kaplan, M. (1981b). Rational Acceptance. *Philosophical Studies*, *40*(2), 129-145.

Kaplan, M. (1995, January). Believing the Improbable. *Philosophical Studies*, *77*(1), 117-146.

Kelly, T. (2010). Peer Disagreement and Higher Order Evidence. In A. I. Goldman & D. Whitcomb (Eds.), *Social Epistemology: Essential Readings* (pp. 183–217). Oxford University Press.

Kerr, N. L., MacCoun, R. J., & Kramer, G. P. (1996). Bias in judgment: Comparing individuals and groups. *Psychological review*, *103*(4), 687.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annu. Rev. Psychol.*, *55*, 623–655.

King, J. C. (2014). Structured Propositions. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 ed.).

King, N. L. (2012, September). Disagreement: What's the Problem? or A Good Peer is Hard to Find. *Philosophy and Phenomenological Research*, *85*(2), 249-272. doi: 10.1111/j.1933-1592.2010.00441.x

King, N. L. (2014, October). Perseverance as an intellectual virtue. *Synthese*, *191*(15), 3501-3523. doi: 10.1007/s11229-014-0418-1

Kinney, D. (2017). Inductive Explanation and Garber–Style Solutions to the Problem of Old Evidence. *Synthese*, 1–15.

Kitcher, P. (1990, January). The Division of Cognitive Labor. *The Journal of Philosophy*, *87*(1), 5-22. doi: 10.2307/2026796

Konek, J., & Levinstein, B. (2016). *The Foundations of Epistemic Decision Theory*. manuscript.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006, February). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology. General*, *135*(1), 36-69. doi: 10.1037/0096-3445.135.1.36

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (p. 117-135). New York, NY, US: Psychology Press.

Lacey, H. (2015, October). 'Holding' and 'endorsing' claims in the course of scientific activities. *Studies in History and Philosophy of Science Part A*, *53*, 89-95. doi: 10.1016/j.shpsa.2015.05.009

Lackey, J. (2008). A Justificationist View of Disagreement's Epistemic Significance. In A. M. A. Haddock & D. Pritchard (Eds.), *Proceedings of the Xxii World Congress of Philosophy* (pp. 145–154). Oxford University Press.

Lasonen-Aarnio, M. (2010). Unreasonable Knowledge. *Philosophical Perspectives*, *24*(1), 1–21.

Lasonen-Aarnio, M. (2013, December). Disagreement and Evidential Attenuation. *Noûs*, *47*(4), 767-794. doi: 10.1111/nous.12050

Lasonen-Aarnio, M. (2014, March). Higher-Order Evidence and the Limits of Defeat. *Philosophy and Phenomenological Research*, *88*(2), 314-345. doi: 10.1111/phpr.12090

Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (Vol. 282). Univ of California Press.

Laudan, R. (1987). The rationality of entertainment and pursuit. In *Rational Changes in Science* (pp. 203–220). Springer.

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, *88*(2), 605–620.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*(3), 177–189.

Leitgeb, H. (2014). The stability theory of belief. *Philosophical Review*, *123*(2), 131–171.

Levi, I. (1974). *Gambling with Truth: An Essay on Induction and the Aims of Science*. Cambridge: The MIT Press.

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. Cambridge, Mass: MIT Press.

Levi, I. (2004a). *Mild contraction: Evaluating loss of information due to loss of belief*. Oxford; Oxford; New York: Clarendon ; Oxford University Press.

Levi, I. (2004b). *Mild Contraction: Evaluating Loss of Information Due to Loss of Belief*. Oxford University Press.

Lewis, D. (1971). Immodest Inductive Methods. *Philosophy of Science*, *38*(1), 54–63.

Lewis, D. (1982). Logic for Equivocators. *Noûs*, *16*(3), 431–441.

Lougheed, K., & Simpson, R. M. (2017). Indirect Epistemic Reasons and Religious Belief. *Religious Studies*, *53*(2), 151–169.

Maher, P. (1993). *Betting on theories*. Cambridge; New York, NY, USA: Cambridge University Press.

Matheson, J. (2015a, April). Are Conciliatory Views of Disagreement Self-Defeating? *Social Epistemology*, *29*(2), 145-159. doi: 10.1080/02691728.2014.907833

Matheson, J. (2015b). Disagreement and Epistemic Peers. *Oxford Handbooks Online*. doi: 10.1093/oxfordhb/9780199935314.013.13

McKaughan, D. (2007). *Toward a Richer Vocabulary for Epistemic Attitudes* (Unpublished doctoral dissertation). University of Notre Dame.

McKaughan, D. (2008). From Ugly Duckling to Swan: C. S. Peirce, Abduction, and the Pursuit of Scientific Theories. *Transactions of the Charles S. Peirce Society*, *44*(3), 446–468.

McMullin, E. (1976). The Fertility of Theory and the Unit for Appraisal in Science. In R. S. Cohen, P. K. Feyerabend, & M. Wartofsky (Eds.), *Essays in Memory of Imre Lakatos* (pp. 395–432). Reidel.

Menzel, C. (2016). Possible Worlds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.).

Mercier, H. (2016, September). The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, *20*(9), 689-700. doi: 10.1016/j.tics.2016.07.001

Mercier, H., & Sperber, D. (2011, April). Argumentation: Its adaptiveness and efficacy. *Behavioral and Brain Sciences*, *34*(2), 94-111. doi: 10.1017/S0140525X10003031

Montmarquet, J. (1992). Epistemic Virtue and Doxastic Responsibility. *American Philosophical Quarterly*, *29*(4), 331-341.

Muldoon, R. (2013). Diversity and the Division of Cognitive Labor. *Philosophy Compass*, *8*(2), 117–125.

Muldoon, R., & Weisberg, M. (2011). Robustness and Idealization in Models of Cognitive Labor. *Synthese*, *183*(2), 161–174.

Nickles, T. (1981). What is a Problem That We May Solve It? *Synthese*, *47*(1), 85–118.

Niiniluoto, I. (1983). Novel Facts and Bayesianism. *British Journal for the Philosophy of Science*, *34*(4), 375–379.

Pagin, P. (2016). Assertion. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 ed.).

Parfit, D., & Broome, J. (1997). Reasons and Motivation. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *71*, 99-146.

Pettigrew, R. (2014, August). *M-Phi: L. A. Paul on transformative experience and decision theory I.*

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press Uk.

Plantinga, A. (1999). "Pluralism: A Defense of Religious Exclusivism&Quot. In K. Meeker & P. Quinn (Eds.), *The Philosophical Challenge of Religious Diversity* (pp. 172–192). New York: Oxford University Press.

Priest, G. (2006). *In Contradiction: A Study of the Transconsistent*. Oxford University Press.

Priest, G., & Berto, F. (2013). Dialetheism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013 ed.).

Priest, G., Tanaka, K., & Weber, Z. (2015). Paraconsistent Logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2015 ed.).

Pritchard, D., Turri, J., & Carter, J. A. (2018). The Value of Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). Metaphysics Research Lab, Stanford University.

Proust, J. (2013). *The Philosophy of Metacognition: Mental Agency and Self-Awareness*. OUP Oxford.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth* (1edition ed.). London ; New York: Routledge.

Rayo, A. (2013). *The Construction of Logical Space*. Oxford University Press.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and instruction*, *11*(3-4), 347–364.

Rinard, S. (2015). No Exception for Belief. *Philosophy and Phenomenological Research*, *91*(2).

Roberts, R. C., & Wood, W. J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Clarendon Press.

Rosenkranz, S., & Schulz, M. (2015). Peer Disagreement: A Call for the Revision of Prior Probabilities. *Dialectica*, *69*(4), 551–586.

Schaffer, J. (2008). Knowledge in the Image of Assertion. *Philosophical Issues*, *18*(1), 1–19.

Schoenfield, M. (2016). An Accuracy Based Approach to Higher Order Evidence. *Philosophy and Phenomenological Research*.

Schwitzgebel, E. (2015). Belief. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 ed.).

Šešelja, D., Kosolosky, L., & Straßer, C. (2012). The Rationality of Scientific Reasoning in the

Context of Pursuit: Drawing Appropriate Distinctions. *Philosophica*, *86*, 51–82.

Šešelja, D., & Straßer, C. (2013). Kuhn and the Question of Pursuit Worthiness. *Topoi*, *32*(1), 9–19.

Šešelja, D., & Straßer, C. (2014). Epistemic Justification in the Context of Pursuit: A Coherentist Approach. *Synthese*, *191*(13), 3111–3141.

Shah, N., & Velleman, J. D. (2005). Doxastic Deliberation. *The Philosophical Review*, *114*(4), 497-534.

Smith, M. (2005). Meta-Ethics. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy* (pp. 3–30). Oxford University Press.

Sosa, E. (1980). The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge. *Midwest Studies In Philosophy*, *5*(1), 3-26. doi: 10.1111/j.1475-4975.1980.tb00394.x

Sosa, E. (2007). *A Virtue Epistemology: Volume I: Apt Belief and Reflective Knowledge*. Oxford: Oxford University Press UK.

Sosa, E. (2009). *Reflective knowledge: Apt belief and reflective knowledge* (Vol. 2). Oxford University Press.

Sosa, E. (2010). The Epistemology of Disagreement. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Social Epistemology*. Oxford University Press.

Sosa, E. (2011). *Knowing full well*. Princeton: Princeton University Press.

Sosa, E. (2015). *Judgment and Agency*. Oxford University Press.

Sprenger, J. (2015). A Novel Solution to the Problem of Old Evidence. *Philosophy of Science*, *82*(3), 383–401.

Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

Stalnaker, R. (1999). *Context and content: Essays on intentionality in speech and thought*. Oxford ; New York: Oxford University Press.

Stalnaker, R. C. (1987). *Inquiry*. Cambridge, Mass.: MIT Press.

Steel, D. (2010). Epistemic Values and the Argument From Inductive Risk. *Philosophy of Science*, *77*(1), 14–34.

Strevens, M. (2003, February). The Role of the Priority Rule in Science. *The Journal of Philosophy*, *100*(2), 55-79.

Strevens, M. (2006). Notes on Bayesian confirmation theory. *manuscript*.

Swanton, C. (2001, October). A Virtue Ethical Account of Right Action. *Ethics*, *112*(1), 32-52. doi: 10.1086/322742

Swanton, C. (2003). *Virtue Ethics: A Pluralistic View*. Oxford University Press.

Talbott, W. (2015). Bayesian Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2015 ed.).

Thoma, J. (2015). The Epistemic Division of Labor Revisited. *Philosophy of Science*, *82*(3), 454–472.

Titelbaum, M. (2015). Rationality's fixed point (or: In defense of right reason). *Oxford studies in epistemology*, *5*, 253–94.

Turri, J. (2017). Experimental Work on the Norms of Assertion. *Philosophy Compass*, *12*(7), e12425.

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141–162). Springer.

Van, P. (2008). Metacognition: Knowing about knowing. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (p. 47-71). New York, NY, US: Psychology Press.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford; New York: Clarendon Press ; Oxford University Press.

Weatherson, B. (2007). Disagreeing About Disagreement. *manuscript*.

Weatherson, B. (2013). Disagreements, Philosophical and Otherwise. In J. Lackey & D. Christensen (Eds.), *The Epistemology of Disagreement: New Essays* (p. 54). Oxford University Press.

Weiner, M. (2007, March). Norms of Assertion. *Philosophy Compass*, *2*(2), 187-195. doi: 10.1111/j.1747-9991.2007.00065.x

Weisberg, J. (2011). Varieties of Bayesianism. In D. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the History of Logic* (Vol. 10). North Holland.

Weisberg, J. (2016). Belief in Psyontology. *The Philosopher's Imprint*, *Forthcoming*.

Weisberg, M., & Muldoon, R. (2009). Epistemic Landscapes and the Division of Cognitive Labor. *Philosophy of Science*, *76*(2), 225–252.

Wenmackers, S., & Romeijn, J.-W. (2016, April). New theory about old evidence. *Synthese*, *193*(4), 1225-1250. doi: 10.1007/s11229-014-0632-x

Whitt, L. A. (1985). *The Promise and Pursuit of Scientific Theories* (Unpublished doctoral dissertation). Dissertation.

Whitt, L. A. (1990). Theory pursuit: Between discovery and acceptance. *PSA: Proceedings of the biennial meeting of the philosophy of science association*, *1*, 467–483.

Whitt, L. A. (1992). Indices of Theory Promise. *Philosophy of Science*, *59*(4), 612–634.

Whittlesea, B. W. A., Jacoby, L. L., & Girard, K. (1990, December). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual

quality. *Journal of Memory and Language*, *29*(6), 716-732. doi: 10.1016/0749-596X(90)90045-2

Williamson, T. (1996, October). Knowing and Asserting. *The Philosophical Review*, *105*(4), 489-523. doi: 10.2307/2998423

Williamson, T. (2000). *Knowledge and its Limits*. Oxford University Press.

Williamson, T. (2008). Why Epistemology Cannot Be Operationalized. In Q. Smith (Ed.), *Epistemology: New Essays*. Oxford University Press.

Winther, R. G. (2016). The Structure of Scientific Theories. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.).

Worsnip, A. (2014). Disagreement About Disagreement? What Disagreement About Disagreement? *Philosophers' Imprint*, *14*.

Yalcin, S. (2015). Figure and ground in logical space.

Yoshioka, S., & Kinoshita, S. (2007). Polarization-sensitive color mixing in the wing of the Madagascan sunset moth. *Optics Express*, *15*(5), 2691. doi: 10.1364/OE.15.002691

Zagzebski, L. (2003). *Intellectual Motivation and the Good of Truth*. Oxford University Press.

Zagzebski, L. T. (1996a). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139174763

Zagzebski, L. T. (1996b). *Virtues of the mind: An inquiry into the nature of virtue and the ethical foundations of knowledge*. Cambridge University Press.

Zollman, K. J. S. (2009). Optimal Publishing Strategies. *Episteme*, *6*(2), 185–199.

Zollman, K. J. S. (2010). The Epistemic Benefit of Transient Diversity. *Erkenntnis*, *72*(1), 17–35.