RATIONALITY WITHOUT REPRESENTATION

by

Alejandro Pérez Carballo

Licence (2003), Maîtrise (2004), Université de Paris 1
Master (2005), Université de Paris 7

Submitted to the Department of Linguistics & Philosophy
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Author .....

Alejandro Pérez Carballo
June 7, 2011

Certified by .....

Robert C. Stalnaker
Laurance S. Rockefeller Professor of Philosophy
Thesis Supervisor

Certified by .....

Stephen Yablo
Professor of Philosophy
Thesis Supervisor

Accepted by .....

Alex Byrne
Professor of Philosophy
Chair of the Committee on Graduate Students

This thesis was typeset using XeTeX.
This typesetting system was developed by
Jonathan Kew based on a merger of e-TeX
—an extension of Donald Knuth's TeX
created by the NTS group—and modern
font technologies.

The body text is set in 12/14.5pt on a 28pc
measure with Minion Pro, an update on a
neohumanist typeface designed by Robert
Slimbach in 1989. Other fonts include Myriad
Pro, a humanist sans-serif typeface designed
by Robert Slimbach and Carol Twombly in 1992,
and Apple's Menlo, a monospaced font based
on Bitstream Vera.

The layout of this document was created using
Peter Wilson's memoir package, based on
recommendations given in Robert Bringhurst's
The Elements of Typographic Style.

Rationality without representation

by

Alejandro Pérez Carballo

Submitted to the Department of Linguistics & Philosophy on June 7, 2011
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
at the Massachusetts Institute of Technology

This dissertation is about whether and how non-representational attitudes could play a role in our theories of rationality. In Chapter 1 ('Negation, expressivism, and intentionality') I argue that the best explanation for why two mental states are inconsistent need not presuppose that such states are representational—that they have, in the jargon, *truth-conditions*. I use this to provide a solution to the 'negation problem' for metaethical expressivism. In Chapter 2 ('Structuring logical space') I sketch an account of mathematical practice along non-representational lines. I show how it can do justice to the applicability of mathematics, and propose ways in which one's epistemic goals can impose substantial constraints on which mathematical theories to accept. Chapter 3 ('Good questions') provides a general account of the way in which rationality constrains changes in our hypothesis space. In particular, I show how some such changes can be better than others by placing the discussion within a general framework of rational dynamics, on which rational epistemic change involves maximizing expected *epistemic* utility.

Thesis supervisor: Robert C. Stalnaker
Title: Laurance S. Rockefeller Professor of Philosophy

Thesis supervisor: Stephen Yablo
Title: Professor of Philosophy

CONTENTS

Abstract 3

Acknowledgments 7

Introduction 10

1 Negation, Expressivism & Intentionality 13
- 1.1 Preliminaries 13
- 1.2 The problem 14
- 1.3 Incompatibility and content 23
- 1.4 Content attribution 25
- 1.5 Inconsistency and truth-conditions 27
- 1.6 Explaining incompatibility 29
- 1.7 Morals 31
- 1.8 Conclusion 32

2 Structuring logical space 33
- 2.1 Preview 35
- 2.2 Mathematics as a source of conceptual resources 36
- 2.3 Accepting a mathematical theory 39
- 2.4 Constraints on mathematical theorizing 46
- 2.5 Deduction 49
- 2.6 Outstanding issues 55
- 2.7 Conclusion 56
- 2.A Appendix: Content and Semantics 58

3 Good questions 67
- 3.1 Evaluating questions 68
- 3.2 Epistemic decision problems 71
- 3.3 The epistemic value of questions 75
- 3.4 Counterfactual resilience and explanation 77
- 3.5 Measuring explanatory potential 82

3.6 Immodesty and Epistemic imagination 85
3.7 Rational dynamics and epistemic value 87
3.8 The rationality of conceptual change 90
3.9 Conclusion 94
3.A Appendix 95

Bibliography 103

ACKNOWLEDGMENTS

I owe far too much to far too many people. It is thus likely—very much so—that I will unwillingly omit a few.

For early guidance and advice, thanks to Denis Bonnay, Susanna Berestovoy, Serge Bozon, Rafael Tomás Caldera, Carlos DiPrisco, Jacques Dubucs, Paul Egré, Jean Gayon, Katarina Livljanić, François Récanati, Daniel Rodríguez Navas, and Stevo Todorčević.

I have been fortunate to be a part of the department of linguistics and philosophy at MIT for the last six years. Faculty, staff, visitors, and fellow graduate students have made this a terrific place to do philosophy. In particular, I would like to thank Mahrad Almotahari, Emmanuel Chemla, Dan Greco, Marie Guillot, Caspar Hare, Valentine Hacquard, Sally Haslanger, Sophie Horowitz, Elisa Mai, Rae Langton, Heather Logue, Susanna Rinard, Miriam Schoenfield, and Paulina Sliwa.

For their advice, friendship, and support, I'm especially indebted to Alex Byrne, Sylvain Bromberger, Danny Fox, Irene Heim, Richard Holton, Dilip Ninan, Paolo Santorio, Kenny Walden, and Seth Yalcin. I have learned much from each of you: probably more than I realize.

For helpful conversations on topics related to this dissertation, thanks to Tom Dougherty, Paul Egré, Nina Emery, Nadeem Hussain, John MacFarlane, Vann McGee, Damien Rochford, Nishi Shah, Elliott Sober, Judy Thomson, and Roger White. Thanks also to audiences at the University of Buenos Aires and the École Normale Supérieure for feedback on the material in Chapter 2.

I have lost track of how much I owe to Bob Stalnaker and Steve Yablo—my thesis supervisors—and to Agustín Rayo—consejero infatigable. During countless hours of conversation, I received far more advice, support, and encouragement from them than I could have hoped for. Every idea in this dissertation was either born in or substantially shaped by my conversations with them.

Thanks to many friends and family, especially to Leopoldo Bello, Rossitza Panova, Luis Pérez Oramas, and Dimitre Vavov, for always

keeping an eye on me.

Thanks to Rae, Richard, Eleanor, and Natalie, for providing me with a home away from home during this last year. I cannot thank you enough.

Thanks to my parents (all three of you) and my sister, for their love, patience, and support.

Above all, thanks to Ekaterina Vavova. I would love to list the reasons why, but that would take us far beyond the scope of this dissertation.

This work is dedicated to the memory of Esther Tropper Cedeño and Manuel Carballo Saavedra, my grandparents.

La sola diferencia es que los filósofos publican
en agradables volúmenes las etapas intermediarias
de su labor y que yo he resuelto perderlas.

Pierre Menard
Bayonne, September 30, 1934

INTRODUCTION

Philosophers are particularly interested in questions like 'what is it to assert that so and so?' or 'what is it to believe that such and such?'. A common strategy has been to tackle these questions in full generality—at least, in as much generality as seems reasonable given basic grammatical facts—by relying on the notion of a *proposition*. To assert that so and so is to put forth the proposition that so and so. To believe that so and so is to stand in the belief relation to the proposition that so and so.

As answers to the starting questions these are hardly satisfactory. We need to say something about what it is to put forth a proposition, what it is to stand in the belief relation to a proposition and, crucially, what a proposition is. Doing that without abandoning the ambition of the project is no easy task. Whatever a proposition turns out to be, it must be play the right role both when accounting for our discourse about medium-sized dry goods and when accounting for our discourse about god, morality, or what have you.

A less ambitious strategy, famously employed by metaethical expressivists, proceeds in a piecemeal fashion. It is open to treating different domains of discourse in importantly different ways. The strategy is to ask, not what asserting a declarative sentence is, but what asserting a declarative *moral* sentence is, or what asserting a declarative sentence about observables is. Similarly, one asks not what it is to have a belief simpliciter, but what it is to have a *moral* belief, etc. Ideally we would want to be able to say something general about all these kinds of discourses. After all, reflection on our discursive practices suggests that they have much in common. But we need not take this to be a nonnegotiable constraint on the project. We should instead start by looking at the role that our moral and mathematical theorizing play in our cognitive and conative lives, and try to make room within a plausible picture of ourselves and the world around us for states that can play those roles.

My conjecture is that this is the best way of making sense of our moral and mathematical practices while holding on to the cluster of

prejudices some of us like to call *naturalism*. For a lot of the difficulties that arise in the metaphysics and epistemology of morality and mathematics, I think, have their roots in this seemingly innocent claim: that our moral and mathematical thought are of a type with our thought about the concrete world.

But I do not have an argument that this is the only way to go. Rather, I wanted to explore the prospects of developing such a strategy. In a way, the question that ties these three papers together is this: if we reject the assumption that our moral and mathematical thought are representational, how can we make for room for them within a plausible theory of rational inquiry?

I do not claim to have an answer to this question. But I hope that what I have to say here gestures in the right direction.

NEGATION, EXPRESSIVISM & INTENTIONALITY

According to metaethical expressivists, when engaged in moral discourse we do not take a stance on what the world is like—we are not in the business of describing the world. In the jargon, expressivists deny that 'moral sentences' have *truth-conditions* (at least in any sense in which truth-conditions have much explanatory leverage). Thus, expressivists cannot appeal to truth-conditions to explain the following basic fact: any sentence involving moral language is inconsistent with its negation.[1]

Some have taken this observation to raise a devastating problem for expressivist (see Schroeder, 2008b, for discussion). The reason is that descriptivists can give a good explanation of inconsistency in terms of the sentences' truth-conditions. And any explanation of inconsistency expressivists can offer will be much worse than the one descriptivists have to offer.

I want to offer a solution to this problem. But I will not offer *another* explanation on behalf of the expressivist and argue that it is as good as the one the descriptivist has to offer. Rather I will argue that the descriptivist cannot offer a good explanation of inconsistency in terms of truth-conditions. Moreover, I will argue, if inconsistency *is* something that needs to be explained, then any plausible explanation available to the descriptivist will be available to the expressivist.

1.1 PRELIMINARIES

Expressivism about moral discourse (e.g. Blackburn, 1998; Gibbard, 1990; Hare, 1952) is a conjunction of two claims. First, that when asserting declarative 'moral sentences' I do not describe a way the world is. Second, that the point of asserting one such sentence is to express a non-cognitive attitude of some kind.

1 Note that it is not obvious how to present the challenge to the expressivist in the first place: if inconsistency *just is* a matter of having incompatible truth-conditions, the expressivist will just deny that moral sentences are inconsistent. I will come back to this below.

An example might help. When I claim that Tom is a cannibal, I am describing a way the world is, viz. it is such that Tom is a cannibal. The point of making this claim is to tell you that the world is that way, to provide you with information. If you take my word for it, you will learn that one way the world could have been—one in which Tom is not a cannibal—is not the way the world in fact is.

According to *descriptivists*, the same is true of my claim that cannibalism is wrong. When I tell you that cannibalism is wrong I am describing a way the world is: a realm of 'queer' properties, perhaps, or our conventions, or what have you. In contrast, expressivists hold that when I claim that cannibalism is wrong I am *not* describing a way the world is. Rather, to a first and rough approximation, I am expressing my *disapproval* of cannibalism. The point of making that claim is to evince my disapproval, and perhaps to get you to disapprove, of cannibalism.

The distinction between expressivism and descriptivism can be drawn in domains other than morality. There are expressivist theories about epistemic modality, epistemic normativity, causality, and probability judgments, to name a few.[2] I think most of the issues raised here do not depend on the specific flavor of the expressivist/descriptivist distinction under consideration. But because of its familiarity, and because it is typically taken to be the *locus* of the negation problem, I will focus mainly on expressivism about moral discourse.

1.2 THE PROBLEM

Consider:

(1) Cannibalism is wrong.

(2) Cannibalism is not wrong.

These two sentences are inconsistent. An adequate account of moral discourse should be able to explain why—this much seems uncontroversial. The worry (cf. Schroeder, 2008b; Unwin, 1999, 2001, *inter alia*) is that expressivists cannot give a satisfactory explanation of this.[3]

2 See, e.g. Yalcin forthcoming, Field 2009, Blackburn 1990, and Price 1983, respectively.

3 Schroeder 2008b claims to provide a solution to the problem. But as he himself goes on to say, it is a consequence of his solution that the semantic program of expressivism

Now, in order to determine whether there is a problem here, we need to flesh out the explanandum so that it is something expressivists and descriptivists can agree on. And this is trickier than it seems.

1.2.1 *The explanandum*

First, one could have a *syntactic* conception of inconsistency. Two sentences are inconsistent, in this sense, just in case one is the result of prefixing the other with a 'not'. But it is hard to see what it would be to *explain* why two sentences are inconsistent, in this sense. Or perhaps one can say that two sentences are inconsistent just in case they license an inference to an explicit contradiction, i.e. to a sentence of the form $\ulcorner \varphi \wedge \neg \varphi \urcorner$.

On this way of understanding inconsistency, what needs to be explained is why (1) and (2) license an inference to a sentence of the form $\ulcorner \varphi \wedge \neg \varphi \urcorner$. But a reasonable explanation of this will have little to do with negation. All we need is an explanation of why the following is a permissible pattern of reasoning:

$$\varphi, \psi \vdash \varphi \wedge \psi.$$

And this could plausibly be explained by the fact that it is constitutive of the meaning of '$\wedge$' that such a pattern of reasoning *is* permissible.

Another option would be to go with a *semantic* notion of inconsistency. Two sentences are inconsistent, in this sense, if they cannot both be true at the same time. But if this is the explanandum, then it is no surprise that the expressivist cannot give an adequate explanation. After all, on a standard way of construing the expressivist's claim, (1) and (2) do not have truth-conditions, so *a fortiori* they are not inconsistent in *this* sense.

Further, according to the descriptivist, one can *explain* the fact that (1) and (2) are inconsistent by pointing to the fact that they have incompatible truth-conditions. But then inconsistency cannot just be a matter of having incompatible truth-conditions, or we would have no explanation in the first place.

The point is simply that we need a characterization of inconsistency that allows the expressivist to recognize that (1) and (2) *are* inconsistent. And this characterization must be such that, *prima facie*, the ex-

cannot succeed (cf. Schroeder, 2008a).

pressivist cannot adequately explain why (1) and (2) are inconsistent. Neither the semantic nor syntactic notions considered thus far can fulfill both these roles. So we need to do better.

I suggest we think of the relation between (1) and (2) that needs to be explained as one essentially involving the notion of *rationality*. Thus, on my view, if there's anything that calls out for explanation is that it is irrational to accept both (1) and (2)—as irrational as it would be to accept both 'Tom is a cannibal' and 'Tom is not a cannibal'. I think this is the only non-question-begging way of cashing out the notion of inconsistency: the only thing an expressivist should reasonably be expected to explain.[4]

Moreover, if mental content is prior to linguistic content—regardless of whether mental content involves, at bottom, something like a 'language of thought'—the question of whether φ and $\neg\varphi$ are inconsistent will ultimately appeal to relations among the *mental states* that are associated with the relevant sentences. Semantic properties of and relations among *sentences*—e.g. logical relations—ultimately need to be explained in terms of properties of and relations among the mental states of users of those sentences. Thus, what needs to be explained is why there is something defective about the combination of the two mental states corresponding to the acceptance of (1) and (2).

Why then is the expressivist especially ill-placed to explain inconsistency in this sense? There are two different worries here, and it is important to keep them apart.

1.2.2 *Unwin's problem*

What is it to accept something like (1)? For the descriptivist, the answer is simple. I accept (1) just in case[5]

(3) I believe that Cannibalism is wrong.

4 There is a question as to what acceptance amounts to for the expressivist. One answer is that I accept *s* just in case I am in the state that I would represent myself as being in when putting forth *s* in conversation. This may not be entirely satisfactory, but for the expressivist program to even get off the ground, something in the vicinity must be. For the purposes of this paper I will assume that expressivists have a satisfactory account of acceptance.

5 Assuming I am a competent English speaker. For ease of exposition, I will drop this qualification in what follows.

Similarly, I accept (2) just in case

(4) I believe that Cannibalism is not wrong.

Two things are worth pointing out here. First, in both (3) and (4), I am being ascribed the same type of attitude—a belief. Second, the attitudes ascribed are beliefs towards incompatible propositions.

On a simple version of expressivism, in contrast, I accept (1) just in case

(5) I disapprove of cannibalism.

And I accept (2) just in case

(6) I approve of cannibalism.

Unlike with (3) and (4), in (5) and (6) I am *not* being ascribed the same attitude with incompatible contents. Indeed, as Unwin (1999) pointed out, (6) simply can't be analyzed as ascribing to me disapproval of *anything*, let alone something incompatible with cannibalism. After all, (6) is importantly different from, say,

(7) I disapprove of not eating human beings.

For I can well approve of cannibalism while also approving of those who do not eat other human beings.[6]

Now, everyone should recognize that in (5) and (6) I am being ascribed incompatible attitudes. It is irrational to disapprove of cannibalism while at the same time approving of cannibalism. So it is not obvious how this is a problem exclusive to expressivists of this variety.

Still, consider what the expressivist will presumably say to explain the inconsistency of 'Tom is a cannibal' and 'Tom is not a cannibal'. Here, the attitudes involved are beliefs. Thus, I accept that Tom is a cannibal just in case I believe that Tom is a cannibal. And I accept that Tom is not a cannibal just in case I believe that Tom is not a cannibal.

6 Similar reasoning shows that (6) cannot be understood as equivalent to

(i) It is not the case that I disapprove of cannibalism.

Perhaps I am undecided as to what to think of cannibalism, and I neither approve *nor* disapprove of it.

So in *this* case, the expressivist will say, inconsistency is explained in terms of the same attitude being held towards incompatible contents. If we want a unified explanation of why a sentence and its negation are inconsistent—among other things so we can explain why accepting (1) and (2) is as bad as accepting 'Tom is a cannibal' and 'Tom is not a cannibal'—then perhaps we will be dissatisfied with this variety of expressivism.

Note that the problem does not arise solely because the objects of the attitude verbs in (6) and (5) are gerunds. Suppose we modified our toy model and said instead that I accept (1) just in case

I require that people do not engage in cannibalism.

There simply is no sentence *s* such that I accept (2) just in case

⌜I require that *s*⌝

is true. For the natural thing to say, on this view, is that I accept (2) just in case

I tolerate that people engage in cannibalism.

and in order to define *tolerating that p* in terms of *requiring*, we need two negations: *requiring that p* is the same as *not tolerating that not p*. So again, we would be unable to analyze the attitudes of accepting (1) and (2) in terms of one attitude with different, incompatible contents.

Having a unified explanation of the different ways in which our attitudes can be inconsistent is not the only reason one might insist that expressivists provide an alternative explanation of the inconsistency of (1) and (2). For it could be that the only way of making sense of inconsistency—in the sense we are interested in—is by looking at attitudes towards incompatible contents. As Schroeder puts it, "[a]ll of the good paradigms that we have of what Gibbard calls disagreement in attitude arise in cases of the same attitude toward inconsistent contents [and this] is the kind of feature to which expressivists are intelligibly entitled to appeal in their explanations" (Schroeder, 2008b, p. 577).

In my view, what this suggests is that this simple version of expressivism is not good enough. For this way of characterizing the relevant attitudes masks the way they logically relate to each other. Fortunately, there are other varieties of expressivism available on the mar-

ket—varieties that recognize that we should characterize acceptance of (1) and acceptance of (2) as attitudes had towards incompatible contents. In particular, one can follow Gibbard (1990) and understand moral judgment in terms of the attitude of *acceptance of a system of norms*.

In its simplest form, the account goes roughly as follows. First, let a *complete* system of norms be a function that assigns a three-fold partition of courses of actions to each possibile world. Intuitively, this partition splits actions according to whether they are *required*, *optional*, or *forbidden* by the system of norms at the given world.[7]

Now, assume that you know what world you are in—assume, that is, that you know all the non-normative facts. On this view, you accept (1) just in case every complete system of norms compatible with what you accept classifies cannibalism as forbidden relative to the world you are in. Similarly, you accept (2) just in case no system of norms compatible with what you accept forbids cannibalism. In other words, you accept (2) just in case every system of norms compatible with what you accept either forbids cannibalism, or deems it optional.[8]

This is only the beginning of the story. In particular, we need to know what it is for a complete system of norms to be compatible with what you accept. But already we are in a position to see that we can analyze accepting (1) and accepting (2) in terms of an attitude (acceptance) had towards incompatible contents, viz. disjoint sets of complete systems of norms. For the set of complete norms that corresponds to your acceptance of (1) is disjoint from the set of complete systems of norms that corresponds to your acceptance of (2). In fact, the set of systems

7 Note that Gibbard's official presentation also allows for complete system of norms to declare an action, in any circumstance, as required, prohibited, or optional (Gibbard, 1990, p. 87f). Thus, *contra* Schroeder 2008b, there is no reason to think that expressivism has a technical problem that makes it unworkable (p. 586). I think Schroeder's complaint is based on the details of the presentation in Gibbard 2003. But there the focus is on judgments of *what to do* in a given decision situation, not on what is morally required. The 'solution' offered in Dreier 2006b to the problem of distinguishing indifference from indecision is just a variant of Gibbard's original account.

8 The assumption of factual omniscience is there only to simplify the exposition. Strictly speaking, on this account we characterize an agent's mental state using *pairs* of the form $\langle w, n \rangle$, where w is a possible world and n is a complete system of norms. An agent characterized by such a set S accepts that cannibalism is wrong just in case in every $\langle w, n \rangle \in S$, cannibalism is forbidden by n in world w.

of norms associated with (1) is the set-theoretic complement to that associated with (2). As a consequence, Unwin's problem, as originally stated, disappears.

Now, perhaps there is a version of Unwin's worry that can be raised even for this version of expressivism. Here is, again, Unwin:

> [T]o rule out the set of completely opinionated credal–normative states in which *s* holds is not to rule out *s*, but to rule out all the maximal ways in which *s* could be accepted: that is to say, it is not *s* itself, but the acceptance of *s* that is excluded. (2001, p. 66)

As I understand it, the complaint is that the attitude characterized by

$$\{n : \text{cannibalism is not forbidden according to } n\}$$

is not equivalent to accepting that cannibalism is not wrong, but rather to that of *not* accepting that cannibalism is wrong. And this distinction, as Unwin rightly points out, is an important one.

But the first of Unwin's claims is just false. When all the norms compatible with what I accept either require cannibalism, or at least deem it optional, I am not merely refusing to accept that cannibalism is wrong. This is so for the same reason that, when in all worlds compatible with what I accept Tom is not tall, I am not merely refusing to accept that Tom is tall. The fact that we allow for an intermediate category—that of it being *optional* to engage in cannibalism—is irrelevant. Suppose in every world compatible with what I believe, Secretariat either did not compete in the 1973 Kentucky Derby or he did compete but he lost. Then it is not just false that I believe that Secretariat won the Derby: rather, I believe that Secretariat did *not* win.

Still, you might wonder whether there is room for the distinction Unwin thinks we ought to make. Grant me that

$$\{n : \text{cannibalism is not forbidden according to } n\}$$

characterizes the attitude of accepting that cannibalism is not wrong. How are we to characterize the attitude of *not* accepting that cannibalism is wrong? That these attitudes are distinct should be clear: it could be that I don't accept that cannibalism is wrong without accepting that cannibalism is not wrong. How should we characterize *that* attitude?

On the face of it, the answer is straightforward: let there be norms compatible with what I accept according to which cannibalism is forbidden, and norms according to which it is not. Then it will not be so that every norm compatible with what I accept makes cannibalism forbidden. But it will neither be that no norm compatible with what I accept prohibits cannibalism.[9]

1.2.3 *The problem of stipulation*

On this way of conceiving of the expressivist framework, we have a certain set of objects (collections of complete systems of norms) that are used to characterize the mental states of moral agents. The view is thus structurally identical to descriptivism: they both use an algebra of sets of points (possible worlds, for the descriptivist; complete systems of norms, for the expressivist) to characterize the relevant mental states. They both read off relations of compatibility and incompatibility between mental states from set-theoretic relations between the sets in the algebra. The difference is that the attitudes characterized by the relevant objects are representational states, in one case, and conative states, in the other.

This leads us to the second problem that is often identified as 'the negation problem'. In explaining why (1) and (2) are inconsistent, the expressivist is offering an analysis in terms of the attitude of *acceptance* of a system of norms. She is suggesting we analyze these attitudes as relations between an agent and a set of complete systems of norms. And she is stipulating that the set assigned to accepting (1) is the comple-

9 There is a lingering problem, one I will not have time to come back to. As Dreier 2006a points out, there is a question of how the attitude characterized by a system of norms all of which characterize cannibalism as optional differs from the attitude characterized by a set of norms that includes some complete systems of norms that characterize cannibalism as required and some that classify it as forbidden. Unwin seems to be sensitive to this problem, when he appears to consider the suggestion I make above only to reject it: "it is unclear how we can accommodate the fact that I might accept not-*s* (as opposed to merely not accepting *s*) in a very provisional and unconfident way." I think there is a genuine question here, viz. how to distinguish settled agnosticism as to whether cannibalism is wrong from acceptance that cannibalism is not wrong (cf. Gibbard, 2003, p. 73). Gibbard's model clearly provides us with the abstract resources to make the distinction, but there is a question of whether the abstract model correctly identifies the two distinct attitudes. I think it does, but this is something beyond the scope of this paper.

ment of the set assigned to accepting (2). But for this to be an explanation of why (1) and (2) are inconsistent, there must be an independent story as to why two attitudes characterized with two disjoint sets of objects cannot be rationally held. Merely stipulating that two attitudes are incompatible just in case they are characterized by two disjoint sets of objects does nothing to explain *why* the attitudes are incompatible.

Indeed, this is at the heart of Schroeder's version of the negation problem for Gibbard's expressivism:

> By assigning the complement set of [complete systems of norms] to a negated sentence, all that Gibbard's account does, is to stipulate that it is to express a state of mind that is inconsistent with the state of mind expressed by the original sentence. But it does nothing to tell us what that state of mind is like, or why it is inconsistent with the state of mind expressed by the original sentence. (Schroeder, 2008b, p. 585).

A similar complaint is voiced by Jamie Dreier:

> Why is there any incoherence in tolerating something and also requiring its contradictory? That is what we are supposed to explain. It's no good just to posit that they are incoherent. The question of how attitudes are logically related, I think, is the same as the question of negation.

The problem, as I understand it, is this. If by 'incompatible contents' we mean 'disjoint sets', then Gibbard has provided us with some reason for thinking that the attitudes of accepting (1) and (2) can be understood as attitudes towards incompatible contents. But then he owes us an explanation for why attitudes that can be characterized with disjoint sets of that kind cannot rationally be held at once. If by 'incompatible contents' he means 'objects such that attitudes characterized with those objects cannot rationally be held at once', then he has provided us with no reason for thinking that the relevant attitudes can be characterized with incompatible contents.

Now, suppose the expressivists convince us that the relevant systems of norms can be used to adequately characterize the attitudes expressed by (1) and (2). Further, suppose they convince us that the attitudes

expressed by (1) and (2)—attitudes characterized using disjoint sets of systems of norms—are rationally incompatible. What else must she do in order to explain why (1) and (2) are inconsistent?

1.3 INCOMPATIBILITY AND CONTENT

There is a difficult question looming in the background. Are facts about what content attitudes have explanatorily prior to facts about what their logical relations are?

Suppose you think that one must explain facts about how mental states logically relate to one another in terms of facts about their content. Then the explanation for why the content of the attitude of accepting (1) is a set disjoint from that of the attitude of accepting (2) must not, on pain of circularity, appeal to the inconsistency of the relevant attitudes.

Let me first point out that while the exposition in Gibbard (2003) suggests that facts about 'disagreement'—about which attitudes 'disagree', or are incompatible, with one another—play a large role in justifying the attribution of content, we have been given no reason to think that this must be so.

Recall the particular version of metaethical expressivism introduced in §1.2.2. I accept that cannibalism is wrong, on that view, just in case every system of norms compatible with what I accept forbids engaging in cannibalism. I accept that cannibalism is not wrong just in case *no* system of norms compatible with what I accept forbids cannibalism. We have thereby specified a particular assignment of content to the attitude of accepting (1) and to the attitude of accepting (2). And, at least *prima facie*, we have not presupposed that accepting (1) and accepting (2) are incompatible. Rather, we have characterized two attitudes—disapproval of cannibalism and tolerance of cannibalism—that we had independent reason for thinking were incompatible.

Perhaps the worry is that, until we are told what it is for a system of norms to be compatible with what I accept, we haven't been told what attitudes is characterized by any system of norms. In other words, until we know what it is for a system of norms to be compatible with what I accept, we won't know whetherthe above two sets of systems of norms characterize attitudes that are rationally incompatible.[10]

10 This might be what Schroeder has in mind, when he writes (2008b, p. 585): "Gibbard's

What would it take to establish that a given abstract object can be used to characterize a particular psychological state? This is a good question, but one we need not answer here. For reflection on what accepting a system of norms *is* is enough to show that the two sets of norms in the previous paragraph characterize incompatible psychological states. I will borrow from the discussion of acceptance in (Gibbard, 1990). The discussion will be highly abstract: Gibbard spends the first two parts of the book painting a picture of what acceptance of a system of norms is, and how it fits within a plausible theory of human psychology—I cannot do justice to that in a few sentences. Nevertheless, I believe what I will say should be enough to illustrate why the two sets of systems of norms above characterize incompatible attitudes.

Start with a simplified story: a complete system of norms is compatible with what an agent accepts just in case in *no* actual or possible circumstance is the agent disposed to act in ways the system of norms classifies as forbidden. A system of norms, on this view, is a partial characterization of a set of dispositions.

This story is too crude for at least two reasons. First, we are sadly well aware of the ways in which our actions can diverge from what we value. It could be that, presented with the opportunity to escape unnoticed after accidentally scratching a car in a parking lot, I would be disposed to take it. And this even though I would consider what I did to be wrong. Second, an ill-informed fellow could act in ways that, by his own lights, are wrong. Perhaps I'm fool enough to believe that the scratch I made will magically vanish after a couple of minutes. Still, had I been told the simple fact that it won't, I would have never chosen to flee.

We can adjust the story in different ways. For example, we can say that a complete system of norms is compatible with what an agent accepts just in case in *no* actual or possible circumstances is the agent disposed to act, without thereafter feeling guilt, in ways that are classified as forbidden according to that system of norms in the circumstances *as she believes them to be.* This may not be quite right—we must acknowledge that accepting that cannibalism is wrong will also affect how I feel

formalism gives us a way of generating complex definite descriptions in order to pick out the states of mind expressed by complex sentences, but no grounds to think that those descriptions actually refer, other than sheer optimism."

about and behave towards others who engage in cannibalism—but it is close enough to be of help for the question we are interested in.

What would it take for an agent to be such that

$$\{n : \text{cannibalism is forbidden according to } n\}$$

correctly characterizes his psychological state? For one, our agent must never be disposed, without a subsequent feeling of guilt, to engage in cannibalism. Thus it would be irrational of her to also feel disposed to engage in cannibalism without a subsequent feeling of guilt. In other words, it would be irrational for her to be in a psychological state that we could characterize with

$$\{n : \text{cannibalism is not forbidden according to } n\}.$$

Perhaps this story relies, implicitly, on the assumption that disapproving of cannibalism and tolerance of cannibalism are incompatible. Perhaps this way of attributing content to the relevant states presupposes, in some other way, what we were supposed to explain. I now want to argue that, even if this is so, the expressivist is no worse off because of this. For *descriptivists*, I will argue, also need to rely on prior relations of incompatibility among mental states in order to assign content to them. To do this, I need to digress briefly to make explicit a few working assumptions.

1.4 CONTENT ATTRIBUTION

Plausibly it is not a brute fact that particular brain states of mine have the content that they do. On a widely shared view, this is something that must ultimately be explained in broadly non-intentional terms. This follows from the assumption that our mental life is part of the natural order: that facts about what beliefs and desires we happen to have are just physical facts. So it must in principle be possible to explain why some of our mental states are beliefs or desires that have the content that they have in broadly physical terms.

What could this explanation look like? Here is a possible candidate. Beliefs and desires are correlative dispositions to act.[11] To believe that

11 Cf. Stalnaker 1984. Incidentally, some expressivists—e.g. Gibbard 1990—seem to explicitly endorse a picture along these lines.

Tom is a cannibal is to be disposed to act in a way that would satisfy one's desires in a world in which Tom is a cannibal and all one's other beliefs are true. Similarly, to desire that cannibalism be outlawed is to be disposed to act in ways that would tend to bring it about that cannibalism is outlawed in worlds in which one's beliefs are true.[12]

The proposition that gets assigned as the content of your belief that Tom is a cannibal is an abstract object that is used to characterize your behavioral dispositions (ditto for your desire that cannibalism be outlawed). I relate to the proposition that Tom is a cannibal in virtue of having the belief that Tom is a cannibal. But this is not some mysterious, non-natural relation, involving some special faculty that gives me access to the realm of abstract objects. Rather, it is much like the way in which a chair can relate to the number 4.2 in virtue of weighing 4.2 pounds.[13] There is no mystery as to how a chair can relate to an abstract real number—there is no need to assume that my chair has some super-natural faculty that allows us to reach out to some platonic realm.

This way of thinking about content attribution—where propositions are used to characterize our mental states much like numbers are used to characterize weights—has much to recommend it. But for our purposes it is enough to point out this much: something like this seems like the only way of solving the problem of intentionality, viz. the problem of explaining why our beliefs and desires have the content that they do. And while one could insist that this is not a problem that needs to be solved, it is only reasonable that the expressivist be allowed to assume otherwise. Recall the status of the dialectic thus far: the expressivist offers a way of assigning disjoint contents to the attitudes expressed by (1) and (2). The descriptivist complains that such an assignment of content is merely a stipulation: she requests an explanation for why those attitudes should be assigned the contents the expressivist assigns to them. It is only reasonable that the descriptivist herself has an explanation of why the belief that cannibalism is wrong gets assigned the content that it does.

Now, the story I've thus far described is overly simplistic and highly schematic. In particular, this approach allows for far too much indeter-

12 I suspect that much of what I have to say can be said even if one has alternative conceptions of what beliefs and desires are. But I want to work with a specific proposal, at least for the sake of concreteness.

13 Cf. the Postscript to Field 1978, in Field 2001.

minacy. Since my beliefs and desires get their content *in tandem*, as it were, by changing the contents of my beliefs and desires appropriately, one can end up with distinct pairs of propositions that can characterized the same set of dispositions to act.[14] Thus, more often than not, versions of this story add an extra layer of complexity.[15] At this level of detail, however, we can already draw some conclusions. And these conclusions, as we will see, call into question the idea that facts about what content my mental states have can play a role in explaining facts about which combinations of states are rational.

1.5 INCONSISTENCY AND TRUTH-CONDITIONS

Consider the following analogy. Suppose you want an explanation for why it is colder in Boston than it is in Canberra. I point out it is 20 degrees in Boston, that it is 85 degrees in Canberra, and that 20 is less than 85. I take it this would not be a satisfying explanation. The reason is that part of what justifies our assignment of numbers to magnitudes has to do with structural similarities between the relations between numbers and the relations between the magnitudes themselves. In other words, we use 20 to index the temperature in Boston and 85 to measure the temperature in Canberra among other things because 20 is smaller than 85 and it is colder in Boston than it is in Canberra.

Similarly, we assign some set-theoretic objects as the contents of some mental states because we think that the relations between those objects are structurally similar to those that obtain between the mental states themselves. If we want an explanation for why two mental states relate to each other in a particular way, pointing out that the corresponding set-theoretic objects relate to one another in the relevant way is not going to help. An explanation of the incompatibility of two mental states in terms of their truth-conditions is no more satisfying than the explanation of why Boston is colder than Canberra in terms of relations between numbers.

Saying that two sentences are inconsistent just in case they have incompatible truth-conditions cannot, therefore, amount to an *explanation*. It is because the mental states associated with the relevant

14 For an example, see Stalnaker 1984, p. 17ff.

15 See the discussion of information-theoretic approaches to intentionality in the next section, for an example of ways of doing this.

sentences are incompatible—where the incompatibility here cannot be explained in terms of their contents—that the sentences get assigned incompatible truth-conditions. So, on pain of circularity, we cannot explain why two sentences are inconsistent in terms of their truth-conditions.

Here is another argument for this conclusion, due in its essentials to Hartry Field.[16] On a wide variety of views about what makes a belief state have the content that it has—often called *information-theoretic* accounts of representation—the notion of *reliable indication* plays a crucial role. A belief state, in addition to being action guiding, is also an indicator of how things stand with the world. My belief that Tom is a cannibal—understood as a particular physical state I am in—has the content that it does partly because, in 'normal conditions', it is caused by the fact that Tom is a cannibal. Thus, if conditions are 'normal', one can take the fact that I believe that Tom is a cannibal to indicate that Tom is, in fact, a cannibal.

A problem for these views, however, is that it is hard to cash out what 'normal conditions' are so that they make the right predictions.[17] It may well be that by maximizing reliability under 'normal conditions', our perceptual beliefs get assigned the truth-conditions that they intuitively have. But it is not obvious that, once we move to more complex beliefs, there is anything like 'normal conditions' under which our beliefs indicate that the world is as they represent them to be.

Take a simple example. Lars is a member of a cult. He believes that Obama is an alien, only because his Leader told him so.[18] Now take an assignment of truth-conditions to Lars's belief states that assigns to his belief that Obama is a spy the set of worlds in which Obama is an alien. Consider a slight variant of this assignment of truth-conditions: one under which his belief that Obama is an alien gets assigned the set of worlds in which *his Leader told him* that Obama is an alien. How could we rule out the latter assignment?

It is hard to see how reliability considerations could do so. In other words, it is hard to see what 'normal conditions' must be like so that Lars' belief that Obama is an alien indicates, under those conditions,

16 Cf. Field 1990.

17 Cf. Dretske 1981; Fodor 1975; Millikan 1984, for attempts at cashing out the notion of 'normal conditions'.

18 Cf. Field 1990, p. 108f.

that Obama is indeed an alien. Field's suggestion is that in order to rule out the latter assignment of truth-conditions, we need to require that our assignments of truth-conditions satisfy some sort of *systematicity* requirement. And crucially, there must be a way of specifying what this sytematicity requirement amounts to independently of what the truth-conditions of the states are—after all, this requirement is needed to determine what those truth-conditions are.

I've been relying on a notion of explanatory priority that is somewhat controversial. You may reasonably have reservations about the thought that facts can be neatly arranged in terms of their explanatory relations. Still, there is a claim in the vicinity of the one I've been making that doesn't require such strong assumptions.

The challenge for the expressivist, recall, was to provide a principled reason for thinking that the attitudes expressed by (1) and (2) had incompatible contents. Moreover, this had to be done without presupposing that the relevant attitudes were incompatible. Given the picture of content attribution described in §1.4, however, it follows that the descriptivist herself cannot meet this challenge. For if content attribution is partly aimed at indexing relations of compatibility and incompatibility between our mental states, it is hard to see how we can attribute truth-conditions to belief states without presupposing something like relations of compatibility and incompatibility between those states.

1.6 EXPLAINING INCOMPATIBILITY

If I am right, we cannot explain why the belief that p is incompatible with the belief that not-p in terms of their truth-conditions. In particular, the descriptivist cannot explain why the belief that cannibalism is wrong is incompatible (inconsistent) with the belief that cannibalism is not wrong in terms of their truth-conditions. So the fact that the expressivist cannot adequately explain why the attitudes expressed by (1) and (2) in terms of their contents cannot be held against her.

Should we insist that relations of incompatibility be explained in other terms? I have no argument that we should—I also have no argument that we should not. Perhaps this *is* something that must ultimately be explained. If so, it *could* turn out that the descriptivist has an alternative explanation for why beliefs with incompatible contents are inconsistent that is not available to the expressivist. That this is far

from obvious should at least shift the burden of proof: pending an argument that such an explanation is forthcoming, it is hard to see why the expressivist has some special problem with negation.

Indeed, I think something stronger holds: *any* plausible explanation of inconsistency that is available to the descriptivist will be available to the expressivist. For expressivists to have a problem of negation, there must be some feature of the relevant mental states, specifiable in non-intentional terms, that meets two conditions. First, it must be a feature the descriptivist attributes to my belief that cannibalism is wrong and that the expressivist does not. Second, it must be crucial for explaining why that attitude is incompatible with my belief that cannibalism is not wrong.

The worry is a deep one. At the heart of the debate between descriptivists and expressivists is the question of whether the attitudes expressed by (1) and (2) are beliefs. It is taken for granted that the default position is that they are: the burden is supposed to be on the expressivist to show that they are not. But our moral attitudes do not seem to share the (non-intentional) features that we take stereotypical beliefs to have. Our moral attitudes, we can assume, do not involve causal covariation with the moral facts: after all, many moral truths are supposed to be necessary. Indeed, given some plausible naturalistic assumptions, moral facts have no causal commerce with creatures like ourselves.[19] Yet it is precisely these features that are taken to explain why beliefs have the truth-conditions that they do—why they are *representational* in the first place.

The burden should be on the descriptivist to show that our moral attitudes must be characterized in the same terms as our beliefs about non-moral matters. Otherwise, it is hard to see what it is about expressivism that gives rise to a problem with negation—a problem that, everyone agrees, is supposed to be exclusive to expressivism.

It may be that, at the end of the day, we will just have to take it is a primitive that the attitudes expressed by (1) and (2) are incompatible. And if that creates a problem with negation, it is a problem we will all have.

19 Admittedly, this is a somewhat controversial position. See Harman 1977; Sturgeon 1988 for discussion.

1.7 MORALS

It is time to draw some morals.

It is often claimed that logic is what grounds the norms of rationality.[20] It is irrational to believe that p while at the same time believing its negation because the two are logically incompatible. But in light of the above, it seems that what contents attitudes have cannot play a role in explaining why some patterns of attitudes are irrational. Thus, logical relations between the contents of our mental states cannot be used to ground norms of rationality.

A second moral to draw from the above is that some well-known arguments for the view that intentions involve belief—what Michael Bratman calls *cognitivism about practical reason*—are undermined, for they presuppose that norms of epistemic rationality can be explained in terms of their truth-conditions.[21] Very roughly, the idea is that in order to explain why there are consistency requirements on *intention*, we must assume that intentions involve beliefs. But the reason this is supposed to help explain why it is irrational to intend to φ while at the same intending to not-φ is that the beliefs that are thereby implicated have incompatible truth-conditions, and that it is irrational to have beliefs with incompatible truth-conditions. If I am right, however, the putative explanandum is no more in need of an explanation than the explanans.

Now, with the negation problem out of the way, it is likely that the expressivist can give an adequate natural semantics for our moral language. This semantics will proceed much like the descriptivist's would: by assigning 'propositions' to sentences in a language. Thus, a third moral to draw from the arguments in this paper is that in assigning propositions one is not thereby committed to taking the relevant attitudes to be belief-like. One can use abstract objects (sometimes the very same abstract objects one could use to characterize beliefs) to characterize attitudes of a very different kind.

In an influential paper, Paul Benacerraf argued that giving a compositional semantics for our mathematical language makes it very difficult to give an adequate epistemology of mathematics (Benacerraf, 1973). But the problem only arises if one presupposes that the attitudes char-

20 See, e.g. Broome 2002.

21 Cf. Wallace 2001.

acterized by a compositional semantics must be beliefs—that they must be explained in terms of some sort of representational relation between those states and mathematicalia. If I am right, we have independent reasons to reject that presupposition. Of course, if we do reject it we need a story as to what the relevant attitudes really are. Telling that story, however, is a task for another day.[22]

1.8 CONCLUSION

There is a cluster of objections to metaethical expressivism that are filed under the heading 'the negation problem'. In essence, the objections amount to a challenge: that of explaining why the attitudes that, according to the expressivist, are expressed by (1) and (2) are rationally incompatible. The challenge gets its force from the thought that descriptivist are better placed to give such an explanation.

I have argued that, to the extent that descriptivists have a satisfactory explanation for why those attitudes are incompatible, the expressivist does too. The argument relied on two main observations. First, that the expressivist, like the descriptivist, can characterize the relevant attitudes within an abstract model in which accepting (1) and accepting (2) get assigned incompatible contents. Second, that one cannot give a full justification for why this characterization is warranted without an independent reason for thinking that the relevant attitudes are incompatible.

I suspect that to determine how to explain why the attitudes expressed by (1) and (2) are incompatible, we will ultimately need to explain, in non-intentional terms, what those attitudes really are. But that explanation will settle the question of whether expressivism or descriptivism is on the right track. Until then, we are free to assume that there is no problem with negation.

22 See the discussion in Chapter 2.

STRUCTURING LOGICAL SPACE

It is natural to assume that mathematics is an attempt to discover and describe facts about mathematical phenomena—much like physics, geology and economics are attempts to discover and describe facts about physical, geological and economic phenomena. But it has proven difficult to say what the mathematical facts are, and to explain how our mathematical practice could reliably get at such facts.[1]

The challenge is particularly pressing if we assume that our mathematical theories are largely correct, and that our epistemic capacities are ultimately to be understood in broadly naturalistic terms. So it is not surprising that each of these assumptions has been denied.

Some think we have no good reason to take our mathematical beliefs to be true. Mathematical theorizing can only tell us what things *would be* like if our mathematical beliefs were broadly correct. If our mathematical practice is somehow helpful in inquiry, it is not because of its success at what it sets out to do. Others think we underestimate our cognitive abilities: perhaps we should conclude that we have a non-natural faculty of 'mathematical intuition' that gives us access to the relevant facts. Others yet think we should hold on to these assumptions and conclude that the success of our mathematical theorizing cannot be explained.[2]

But one assumption remains unchallenged: that we have mathematical beliefs (or at least belief-like attitudes of some kind—supposition, make-believe, or what have you.[3]) Few would deny, in other words, that mathematics is an attempt to discover and describe facts of some kind. My goal in this paper is to provide a principled way of denying just that.

1 Cf. Benacerraf 1973; Field 1982.

2 Cf. Field 1980, Gödel 1947, and Burgess and Rosen 1997, p. 45, respectively. Some varieties of fictionalism (e.g. Yablo 2001) don't fall into either of these categories, but they maintain that mathematical thought should be analyzed in terms of belief-like attitudes.

3 One exception seems to be Bishop Berkeley—cf. Berkeley 1732, VII, §14, as well as the *Treatise*, §20—and perhaps David Hilbert, on some interpretations. For discussion, see Detlefsen 2005.

To accept a mathematical theory, on my view, is not to have a belief about some subject matter—at least not if we think of beliefs as essentially attempts to describe some realm of facts. The point of mathematical practice is not to gather a distinctive kind of information—'mathematical information'. It is rather to *structure logical space* in an epistemically useful way. When I accept a mathematical theory, I do not change my view on what the world is like—I do not, to use a familiar metaphor, rule out a way the world could be.[4] Instead, I adopt conceptual resources that allow me to make distinctions between ways things could be—to structure the space of possibilities in ways conducive to discovering and understanding what the world is like.

My suggestion is similar in spirit to some *non-cognitivist* views in metaethics. On these views, to judge that cannibalism is wrong is not to take a stance on what the world is like. Morality is not about getting at the moral facts—rather, it is about how to live, what to do. Similarly, on my view, to make a mathematical judgment is not to take a stance on what the world is like. Mathematics is not about getting at the mathematical facts—rather, it is about how to structure the space of hypotheses with which we theorize about the world.

Non-cognitivists often offer their view as the best way of making sense of the motivating force of moral judgments. But non-cognitivism can also be seen as a way of dissolving the well-known difficulties of accounting for our moral practice. As with mathematics, these difficulties arise around what is perhaps the central question in moral epistemology: how do we come to know moral facts? The non-cognitivists deny a presuppositions of the question, viz. that our moral practice should be understood as involving a relation between ourselves and a realm of moral facts.

I seek to reject a similar presupposition in the metaphysics and epistemology of mathematics. On my view, what needs to be explained is not how we can relate to some realm of mathematical facts, nor how our mathematical practice can reliably reflect what goes on in a far away realm. What we need is an explanation of what we *do* when we do mathematics. We need an account of the goals of our mathematical practice

4 Some think that we cannot fully characterize our cognitive lives with sets of possible worlds alone—e.g. Soames 1987. I agree. My proposal is an attempt to go beyond the possible-world model in order to give a better picture of our mathematical thought. Cf. fn. 11.

that does not make it a mystery how creatures with our interests and abilities could successfully engage in it.

My view is thus a form of *nonfactualism* about mathematics. Our mathematical theorizing does not aim to discover a particular sort of fact. It is, nevertheless, rationally constrained. These constraints don't arise out of some putative domain of facts that we are trying to track. They arise instead from our more general epistemic goals. On my view, we should seek mathematical theories that allow us to isolate information about the physical world that is most conducive to our knowledge and understanding.

2.1 PREVIEW

There are two main aspects to our mathematical practice: deducing new claims and accepting new theories. Most of everyday mathematics involves deducing new claims from previously accepted ones. But when we set forth the axioms of our theory of arithmetic, we did not deduce them from something else. Nor was deduction what led to the 'discovery' of real numbers, or of permutation groups. We simply took up some new mathematical structure as an object of study. We *accepted a new theory* about this structure.

I want to start by focusing on what we do when we accept a new mathematical theory. I will not discuss deduction until §2.5. Of course, a full account of mathematical thought must explain the cognitive accomplishment involved in proving a particular theorem (e.g. that there are infinitely many primes,or that every set is smaller than its power set). But it would be a mistake to try to do so in isolation. After all, talk of 'discovering that there are infinitely many primes' only makes sense against the background of a large body of arithmetical assumptions—a mathematical theory. We first need an account of what it is to accept a mathematical theory before we can say what it is to draw a logical consequence from that theory.

Here is what I will do. I will begin (§2.2) by isolating an important role that mathematical theorizing plays in our cognitive economy. I will use that in §2.3 to build an account of what it is to accept a mathematical theory. I will show how the account differs from one on which we have mathematical beliefs in the ordinary sense. This will lead to an account of the cognitive utility of mathematics, and of how rationality constrains

our mathematical theorizing even if we are not aiming to track some putative domain of facts (§2.4). I will turn to the question of deductive reasoning in §2.5. Before concluding, I will list what I take to be the most pressing outstanding issues (§2.6).

My goal here is to sketch an alternative to factualist accounts of mathematics. I won't be arguing against factualist accounts directly. In part, this is because that would take us too far afield. But more importantly, this is because we will be better placed to make a choice between factualism and nonfactualism only once the nonfactualist alternative is on the table. To my knowledge, no such alternative has been developed in any detail. I hope to change that here.

2.2 MATHEMATICS AS A SOURCE OF CONCEPTUAL RESOURCES

What effect does accepting a new mathematical theory have on our cognitive lives? How is this reflected in our overall mental state?

Here is one uncontroversial answer: when we accept a new mathematical theory, we gain conceptual resources. We gain the ability to articulate propositions about the concrete world that we would be unable to articulate otherwise.

Consider Newtonian mechanics and the discovery of the calculus, or Quantum mechanics and the discovery of Hilbert spaces. In each case, non-trivial amounts of mathematics are necessary to formulate crucial aspects of the relevant physical theories—theories that make claims about what the world is like.

Here is a simpler example:

(1) The number of houses on Elm St is odd.

Whatever your views on number talk, you should agree that (1) tells us something about the concrete world. It is something that would be true if any of the following were true:

> There is exactly one house on Elm St.
>
> There are exactly three houses on Elm St.
>
> There are exactly five houses on Elm St.
>
> ...

What (1) entails about the concrete world might be summarized by the

infinite disjunction of all of such claims. Of course, we do not (and could not) have an infinitary language. We are finite beings after all. But with a little bit of mathematics we are able to learn that there is an odd number of houses on Elm St.[5]

Here is a more interesting example.[6] One of Leonhard Euler's most well-known achievements was the solution of the *Königsberg Bridges problem*:

(KB) Is it possible to tour the city of Königsberg (see Fig. 2.1) crossing each of its seven bridges exactly once, and ending at the starting place?

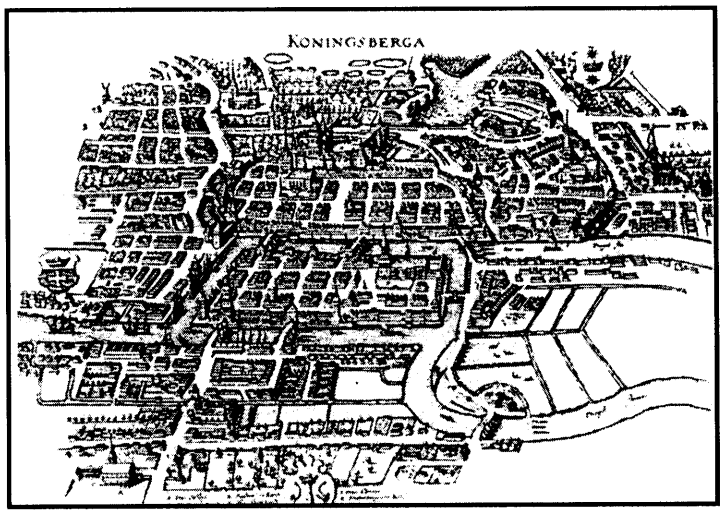


Figure 2.1: Königsberg ca. 1652.

We can reconstruct his solution in two steps. First, he isolated a proposition about the city—call it *Euler's proposition*. Once understood, this proposition can easily be seen to be true. Second, he proved that Euler's proposition entailed that the answer to (KB) is *no*.

5 Some nominalists will object to this—see, e.g. Field 1980. They will insist that our mathematical talk is merely shorthand: we could, if we worked hard enough, express everything we need to express about the concrete world in a finitary non-mathematical language. See Burgess and Rosen 1997 for discussion of the limitations of these reconstructive programs.

6 Cf. Pincock 2007.

I want to focus on the first step. (I will turn to the second step in §2.5.) It is a nice illustration of the way in which new mathematical theories improve our conceptual resources.

What is Euler's proposition? Let me introduce a bit of terminology. Think of a *graph* as a collection of points, or *vertices*, connected to each other by one or more *edges* (see Figure 2.2 for an example).[7] A *path* in a graph is a sequence of vertices and edges, where each edge is between its two vertices. Call a path containing every edge in the graph exactly once an *Euler path*. An *Euler tour* is an Euler path that starts and ends with the same vertex.

Euler's first insight was that the solution to (KB) depends essentially on whether there is an Euler tour in the graph in Fig. 2.2 (where the edges represent the bridges, and the vertices the landmasses). He then

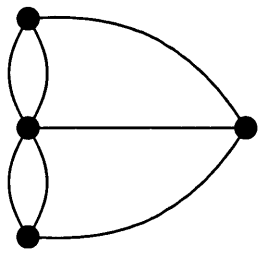


Figure 2.2: A graph representing the structure of Königsberg.

proceeded to give a proof of *Euler's theorem*: that a graph contains an Euler tour if and only if each of its vertices is of even *valence*, where a vertex is of even (resp. odd) *valence* iff it is reached by an even (resp. odd) number of edges.

Introducing this small amount of graph theory allowed Euler to isolate a true proposition ('Euler's proposition') entirely about the bridges of Königsberg:

(EP) The structure of the bridges of Königsberg is a graph at least one of whose vertices is of odd valence.

7 Formally, we can identify a graph with an ordered triple $\langle V, E, f \rangle$, where V and E are the sets of vertices and edges (respectively), and f is a function assigning to each $e \in E$ a two-membered subset of V, so that $f(e)$ is the set of e's vertices.

As you can see by looking at Fig. 2.2, every vertex in the graph representing Königsberg and its bridges is of *odd* valence. In short, (EP) is true. But this proposition is not just one more truth about the bridges of Königsberg. It is one whose connections to other propositions about the bridges are made apparent because of how it is embedded in the theory of graphs. In particular, given Euler's theorem, it follows from (EP) that the answer to (KB) is *no*. So we have a solution to the Königsberg Bridges problem.[8]

2.3 ACCEPTING A MATHEMATICAL THEORY

Accepting a mathematical theory can provide us with new conceptual resources. But how? In particular, how can it provide us with *fruitful* conceptual resources?[9]

I suppose the simplest answer is this: because to accept a mathematical theory *just is* to adopt certain conceptual resources. I will now elaborate on this simple answer to give an account of what it is to accept a mathematical theory. But before doing that, we need to answer a preliminary question: what is it to adopt some conceptual resources? On my view, to adopt new conceptual resources is to make new distinctions among possibilities. Let me explain.

Following Lewis, we can think of the collection of all possibilities as a 'logical' space. A believer, on the Lewisian metaphor, is a traveler trying to locate herself in logical space.[10] So we can think of an agent's belief state as a particular type of map: possibilities compatible with what she believes are spread out all over it. Her goal is to find the point on the map where she is located. When our agent finds out that p, she rules out all those possibilities in which it is not true that p. She thus comes closer to isolating the point on the map corresponding to the way

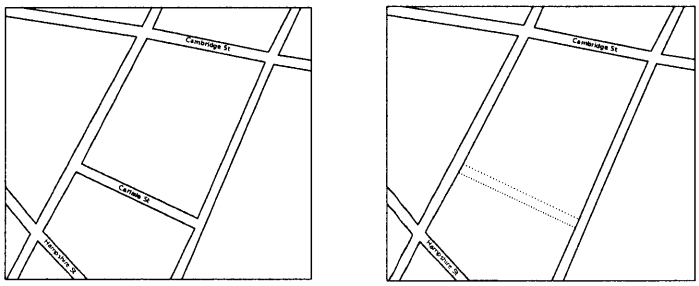
8 I am assuming that (EP) is entirely about the bridges of Königsberg: it simply claims that they are arranged in a particular way (so that there is an isomorphism from the graph in Fig. 2.2 to the city of Königsberg). But nothing hinges on this. If you think (EP) is not entirely about the concrete world, let 'Euler's proposition' refer to the strongest proposition about the concrete world that (EP) entails.

9 This, in a nutshell, is a version of the problem of accounting for the applicability of mathematics. See Steiner 1998, 2005 for discussion.

10 See, e.g. Lewis 1979a. This metaphor can probably be traced at least back to Wittgenstein, but it is most clearly associated with F. P. Ramsey—see Ramsey 1931.

things are.[11]

Some maps are more fine-grained than others. Consider a map that leaves out small streets, like Carlisle St (see Figure 2.3). Using that map alone, a traveler cannot locate herself to the North of Carlisle St, or to the South of Carlisle St. In other words, the agent cannot use the map to demarcate the region that is North of Carlisle St but South of Cambridge St, from the one that is South of Carlisle St but North of Hampshire St.



(a) With Carlisle St. (b) Without Carlisle St.

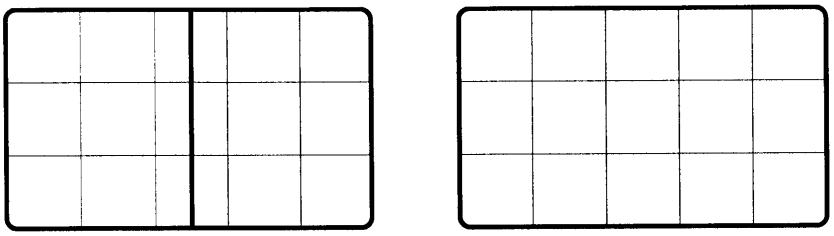
Figure 2.3: Maps with slightly different levels of granularity.

Likewise with beliefs. We can imagine an agent that cannot locate herself exactly in the region of logical space in which quarks are tiny, perhaps because she has never even heard of quarks before. She is thus unable to wonder whether quarks are tiny: she lacks the conceptual resources to distinguish worlds in which quarks are tiny from those in which they're not. It is only when she acquires the ability to make this distinction—the ability to entertain the proposition that quarks are tiny—that her map of logical space can go from the one in 2.4(b) to the one in 2.4(a).[12]

On this way of thinking, an agent's ability to draw on some conceptual resources can be identified with her ability to make new distinc-

11 I am assuming that beliefs should be analyzed in terms of epistemic possibilities. I think of these as metaphysically possible worlds, but nothing in what I will say hinges on this.

12 Cf. Leuenberger 2004 for this way of thinking about entertainability. For related discussion, see Swanson 2006; Yalcin 2008 and the discussion of 'digital' and 'analog' representation in Dretske 1981.



(a) A map that makes the distinction (b) A map that does not

Figure 2.4: Logical space divided by the proposition that quarks are tiny. Think of each point inside the two rectangles as a possible world. The lines correspond to distinctions between those worlds that are made by the map. The worlds in the right half of each rectangle are those in which quarks are tiny.

tions among possibilities. And we can think of the distinctions she is able to make as the propositions she is able to entertain.[13]

Now, if Alice has not heard of quarks, her map of logical space will not make distinctions that depend on how things stand with quarks. But even if she comes to acquire the relevant concepts—even after she acquires the ability to make the distinctions—she may not actually make them: she may not include those propositions among those she takes to be worth gathering evidence for. How things stand with quarks may have no bearing on any question she cares about. So we can represent an agent's conceptual resources by the degree of granularity she is *able* to give to her map of logical space. Which level of granularity she will give to her *working picture of logical space*—those propositions she takes to be worth gathering evidence for, the space of hypotheses that she appeals to for theorizing about the world—will depend on whether she thinks those distinctions are worth making.

I can now give a more fleshed-out formulation of my proposal. Un-

13 It might seem odd to identify conceptual resources with the ability to entertain some propositions. Although I cannot argue for this here, one can capture a lot of our talk of concepts in an apparatus that starts with propositions rather than concepts as its basic component. Very roughly, the idea is to think of concept possession as a closure condition that mimics Evans' (1982) *generality constraint*: to have the concept F just is a matter of being such that if one is able to entertain the thought that x *is* F, one is therefore able to entertain the thought that y *is* F for any y one is acquainted with.

derstanding a mathematical theory can increase an agent's conceptual resources. In coming to understand a mathematical theory, one acquires the ability to entertain some propositions. In coming to *accept* a particular mathematical theory, one comes to adopt the distinctions given by those propositions for the purposes of theorizing. To accept a new mathematical theory is thus to increase the granularity of one's working picture of logical space. Unlike coming to have a belief, accepting a mathematical theory does not involve eliminating any possibilities. Rather, it involves making new distinctions among possibilities.

Now, when I first heard about possums, I acquired the ability to make new distinctions between possibilities—e.g. to distinguish between possibilities in which possums are pests from those in which they are not. Clearly, such a change in my picture of logical space does not involve accepting a new mathematical theory. So we need a principled way of distinguishing the addition of propositions about possums from those additions that do correspond to adopting a new mathematical theory.

Here is a natural suggestion. Mathematics allows us to isolate *structural* features of physical systems. Mathematics gives us ways of carving up logical space where worlds sharing a given structural feature are treated as equivalent. Let me call such propositions *structural propositions*.

It is a difficult question—for reasons that are independent of my view—what a structural feature is. It is thus equally difficult to give an account of structural propositions— structural propositions are those whose truth supervenes on structural features of the world. I cannot provide such an account here. But *very* roughly, a proposition is a structural proposition if its truth depends on the way in which the relevant objects and their parts relate to one another, and not on the identity of the objects themselves.

Some examples might help. An object's shape is a structural feature. A proposition about the spatial arrangement of some objects is arguably a structural proposition. In contrast, the proposition that there are possums in New Zealand is not a structural proposition. For it may be false in a possum-less world in which creatures that are functionally indistinguishable from possums are rampant in New Zealand.[14]

14 Perhaps more controversially, one might think that propositions whose truth is sen-

Recall the Königsberg example from §2.2. Euler's proposition is a structural proposition. Its truth does not depend on features of the city of Königsberg other than the relationships between the bridges and the landmasses they connect. In particular, it does not depend on which materials the bridges are made of, nor on which individuals inhabit the city.

I can now contrast my proposal with other, factualist ones. All parties would agree that there was a change in Euler's cognitive state when he discovered graph-theoretic structures. He was now able to see the bridges of Königsberg as instantiating a particular graph-theoretic structure. (Figure 2.5 is a model of this change.[15])

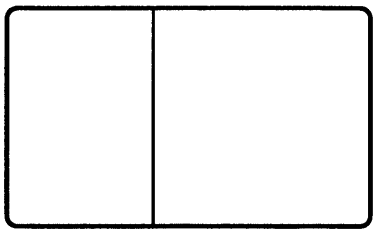
According to the realist, there are some (epistemic) possibilities that Euler ruled out when he discovered graph theory. (If the realist thinks that graphs exist necessarily, she will say those possibilities are 'metaphysically impossible'.) On her view, the change depicted in Fig. 2.5 was not immediate: it involved first eliminating those possibilities and *then* using the newly acquired beliefs to see the city of Königsberg as instantiating the graph in Fig. 2.2. (A realist need not think that these changes take place 'one at a time'. The point is that, on her picture, we can conceptually pull them apart.)

The fictionalist, on the other hand, would perhaps insist that Euler first learned something about some non-actualized possibilities—in which, contrary to fact, our mathematical theories are true. Still, possibilities are being ruled out according to whether they make true counterfactuals of the form 'if the fiction were fact, then p'. Euler then used

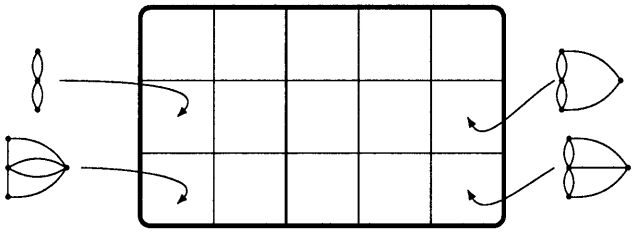
---

sitive to which categorical properties an object has do not count as a structural propositions. See Chalmers 2003; Lewis 2009. Proponents of this view often go on to claim that science only tells us about structural properties of objects. If this turned out to be true, it would follow from my view that the relevant propositions require accepting a substantial amount of mathematics. But that may well be right, given the pervasive role that mathematics plays in the natural sciences.

15 Let me flag something. Figure 2.5 suggests that the change in Euler's cognitive state involves making finer distinctions among possibilities. But it seems intuitively clear that one thing gained by the introduction of graph theory was the ability to see possible configurations of the city as having something in common—to abstract away from details of the city. Thus, it might be natural to think of Euler's change as involving some sort of *coarsening* of logical space. Strictly speaking, as a matter of algebraic fact, any addition of a new proposition to one's picture of logical space will be the result of a refinement, even if the epistemic benefit comes from the induced coarsening. See §2.5 for further discussion.



(a) Possible configurations of the city, divided by the proposition that the answer to (KB) is 'no'.



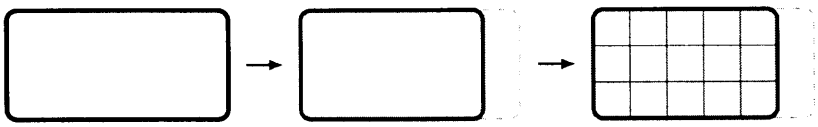
(b) Possibilities further classified according to the graph-theoretic properties of the bridges. Those in the same cell agree on what the bridges' structure is.

Figure 2.5: Two stages in the overall change of Euler's cognitive state. After discovering graphs, Euler acquired the ability to classify possibilities as in (b). Only after being able to classify possibilities this way did Euler gain the ability to entertain Euler's proposition.

these newly acquired beliefs in order to see the city of Königsberg as having the structure of the graph in Fig. 2.2.

Both the realist's and the fictionalist's accounts are thus versions of a *three-stage view*. In contrast, on my proposal the change corresponding to the newly acquired beliefs about graphs does not involve ruling out possibilities of any kind. Rather, it just involves undergoing the change depicted in Fig. 2.5. (See Fig. 2.6 for a contrast between simplified versions of the story according to each of these views.)

The distinctions corresponding to differences in the graph-theoretic structure of the bridges of Königsberg are not the only ones that can be made with the introduction of graph theory. We can ascribe a particular graph-theoretic structure to the bridges of *any* city we are familiar with. To understand the theory of graphs is to be able to put a variety of



(a) A three stage view.

(b) A two stage view

Figure 2.6: Two contrasting accounts of the change in Euler's cognitive state. Here, the grayed out areas correspond to possibilities that have been ruled out. On factualist models, Euler's change occurs in three stages. He first proceeds to eliminate some possibilities (those in which there are no graphs, perhaps), and then puts those new beliefs to use in classifying the different configurations of the bridges. On a nonfactualist model, the change does not involve eliminating any possibilities.

related distinctions to use in one's epistemic endeavors, and to be able to draw connections between those distinctions.

Say that a physical system lends itself to graph-theoretic interpretation if it can be seen as a (partial) interpretation—in the standard, model-theoretic sense—of the axioms of graph-theory. In discovering the theory of graphs, Euler acquired the ability to ascribe graph-theoretic structure to any physical system that lends itself to a graph-theoretic interpretation. Less precisely, though more vividly: Euler acquired the ability to *see* physical systems as graph-theoretic structures.

This observation can be generalized to many mathematical theories. Consider the gain in conceptual resources when an agent manages to see physical systems as interpretations of our language of arithmetic:[16] to take a simple example, she acquired the ability to see any group of three pebbles as having something in common with all and only all groups of $2n + 1$ pebbles for any n (viz. as being a *odd* number of pebbles).[17] Similarly, consider the abilities an agent gains by understanding

16 Note that for this to happen they need not have had anything like *our* language of arithmetic.

17 More generally, for any mathematical structure and any physical system, we can ask

the calculus: e.g., the ability to see different sets of data points as being instances of the same function, or as being all generated by polynomials; the ability to place the claim that a system evolves in a continuous manner within a larger network of relevant claims about theoretically interesting properties of the system.

But we need not assume that mathematical theories were developed with applications in mind. For acquiring the ability to *see* the bridges of Königsberg as a graph-theoretic structure can be done without having this (or any other) application in mind. Very roughly, accepting a mathematical theory is tantamount to acquiring the ability to apply that theory, whether or not one goes on to do so.

The natural question to ask is whether this proposal can be generalized to all mathematical theories. This is no doubt a difficult question. A full answer is beyond the scope of this paper. But in §2.6 I will briefly sketch what I take to be a promising path.

2.4 CONSTRAINTS ON MATHEMATICAL THEORIZING

Our theorizing about the concrete world is constrained by the facts—by what the world is like. I have proposed that our mathematical theorizing does not involve involving a relation between ourselves and a realm of mathematical facts. What then constraints our mathematical theorizing? How can we evaluate mathematical theories from an *epistemic* point of view?

On my account, to accept a mathematical theory is to modify one's working hypothesis space—by making new distinctions, or by abstracting away from others. In other words, it is to take on a particular way of carving up logical space for theorizing about the world. Given that good inquiry is partly a matter of formulating the right hypotheses, one's *epistemic* goals must constrain which way of carving up possibilities one should adopt—and thus, which mathematical theories one should adopt.[18]

---

whether there is a (partial) isomorphism from the mathematical structure to the physical system. The propositions generated by M will be the closure under Boolean operations of all propositions of the form 'S is a physical system isomorphic to M'.

18 Cf. Bromberger 1966, 1988. Frege makes vivid the importance of drawing new boundaries in inquiry: "The more fruitful type of definition is a matter of *drawing boundary lines that were not previously given at all.* What we shall be able to infer from it, cannot be inspected in advance; here we are not simply taking out of the box

We can say more. Which propositions make up one's picture of logical space will constrain which beliefs one will come to have. While different yet compatible sets of beliefs may not differ in how many truths they contain, they may differ in other epistemically significant ways. In particular, which propositions are included in one's system of beliefs will constrain what kind of explanations one can provide. This is not a matter of how many truths a system of beliefs contains: two systems of beliefs that are equally accurate may differ in the type of explanations they can provide.

Take Putnam's famous example.[19] Alice has a small board in front of her with two holes on it, *A* and *B*. *A* is a circle one inch in diameter; *B* is a square one inch in height. She tries to get a cubical peg slightly less than one inch high to go through each hole. Alice believes that the peg can pass through hole *B*, but not through hole *A*.

Compare two different systems of beliefs that could be Alice's. One contains a true description of the microphysical structure of the system consisting of the board and the peg. It also contains a list of the laws of particle mechanics. These, we can suppose, entail that given the microphysical structure of the system, the peg cannot pass through hole *A*, but can pass through hole *B*.

The second system of beliefs takes no stance on what the microphysical structure of the system is. However, it contains the proposition that the peg is cube-shaped, the proposition that hole *A* is round, and the proposition that hole *B* is square. These three (true) propositions in turn entail that the peg cannot pass through hole *A*, but can pass through hole *B*.

We can assume that the two systems of beliefs do not differ in any other significant respect. In particular, they do not differ in how many true propositions they contain. Yet we know enough to see that the second system of beliefs is superior to the first in at least one respect: it allows for a better explanation of why the peg can pass through hole *B*, but not through hole *A*.

This is not to say that the second system of beliefs is better than the first *all things considered*. But it should be uncontroversial that there is

---

again what we have just put in. The conclusions we draw from it *extend our knowledge* [...]." (Frege, 1884, §88, my emphasis)

19 Putnam 1975.

one *pro tanto* reason for preferring it, epistemically, to the first. When evaluating systems of beliefs, we need to look not only at the accuracy of each system, but also at *which* propositions the system includes.

We can return to Euler's example to illustrate this point as well: the addition of graph-theoretic resources is a non-trivial cognitive achievement, one that led to an epistemic improvement in Euler's system of beliefs. Euler used Euler's proposition to explain why one cannot tour the city of Königsberg by crossing each of its bridges exactly once. Now, he could have in principle explained this in terms of the microphysical structure of the city of Königsberg. But even for someone with the computational resources to understand that explanation, the one in terms of Euler's proposition is better for two reasons.[20]

First, the explanation in terms of Euler's proposition is more general: it can apply to a wider variety of cases. General explanations tend to be more satisfactory and thus can be expected to have a high explanatory value.[21] This is why explaining my opening the door by appealing to the claim that someone knocked on it seems more satisfactory than the explanation that appeals to the fact that Tom knocked on it.

Second, the explanation in terms of Euler's proposition manages to abstract away from prima facie *irrelevant* features of the city of Königsberg. *Ex ante*, we are inclined to think that small changes in the microphysical structure of the city of Königsberg would not affect the answer to (KB). Explanations that rely on what we take to be inessential details are worse than those that do. (This is why appealing to beliefs and desires to explain my behavior can be more satisfying than giving a full account of my brain state.[22])

The explanation in terms of Euler's proposition has these virtues because Euler's proposition is a structural proposition. We can expect explanations in terms of structural propositions to have these explanatory virtues. So if what we are after is an increase in valuable explana-

20 More carefully: it is a better explanation for the particular explanatory task at hand. But the two features of the explanation in terms of Euler's propositions that I will go on to discuss tend to make for good explanations more generally, at least given the kinds of things we want to explain.

21 Highly disjunctive explanations may be the exception—so not *any* way of weakening the explanans leads to good explanations. It is not clear why.

22 Cf. Jackson and Pettit 1988. See also Garfinkel 1981; Strevens 2004, and the discussion of *stability* in White 2005.

tory resources, it is worth taking on the expansions that correspond to accepting a mathematical theory.

Now, this does not give us a story about why we have accepted the mathematical theories we actually have. But it shows how one can have principled ways of evaluating distinct expansions of one's picture of logical space, and how these can be seen as arising out of our epistemic goals.[23] More importantly, it gives us the beginnings of an explanation of how creatures with goals and interests like our own could have developed something like our mathematical practice. To have something like our mathematical practice is essentially to have a picture of logical space rich in structural propositions. If our picture of logical space evolved partly by trying to acquire propositions with high explanatory value, it is not surprising that we came to have something like our mathematical practice.

Increasing explanatory resources is not the only goal that our mathematical theories can help us meet. They also allow us to systematize data and make predictions that would be obscured by irrelevant details. More generally, mathematical theorizing can sometimes provide us with helpful computational resources. The example of the bridges of Königsberg shows that much. The discovery of graph theory was crucial for providing an explanation for why the answer to (KB) is what it is. But it was also crucial for proving *that* the answer to (KB) was *no*. To understand how accepting a mathematical theory can play such a role, we need to say something about the role of deductive reasoning in mathematics.

2.5 DEDUCTION

Suppose I am right that to accept a mathematical theory is primarily to add structural propositions to one's working picture of logical space. How does deductive reasoning work on this picture? In particular, what is it to accept a logical consequence of a theory one accepts?

23 In Chapter 3 I examine this question in more detail. By placing the discussion within a general framework of rational dynamics—on which rational epistemic change involves maximizing expected *epistemic* utility—I argue that one can make sense of expansions that are epistemically rational. The key claim is that expansions can lead to epistemic states that are more stable, and that epistemic utility maximizers seek to increase the stability of their epistemic states.

To answer these questions, I will first introduce an abstract framework for thinking about deductive reasoning for factual beliefs.[24] I will build on this framework to sketch an account of deductive reasoning for mathematical thought.

More often than not, our beliefs are not deductively closed. David Lewis tells the story of how he 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."[25] While Lewis did believe all these things, it would be a stretch to say that he believed their conjunction—an obvious inconsistency in light of his background beliefs.

Lewis' proposal was to think of his belief corpus as compartmentalized: rather than thinking of his actions as governed by one inconsistent body of beliefs, we should think of them as governed by distinct bodies of beliefs in different contexts. In some contexts, his actions were guided by the belief that Nassau St (and the railroad) ran roughly north-south. Perhaps when asked where north was, while on Nassau St, he would point in a direction parallel to it. In other contexts, his actions were guided by the belief that the railroad ran roughly east-west. Perhaps when asked where north was, while on the train, he would point in a direction perpendicular to the tracks.

The moral is that we should think of all agents who appear to have inconsistent beliefs as having distinct consistent fragments that are incompatible with each other.[26]

On this view, failures of logical omniscience are due to fragmentation. Lewis believes that Nassaut St runs roughly north-south, that Nassau St and the railroad run roughly parallel to each other, but he fails to believe that the railroad runs roughly north-south. He has two fragmented bodies of belief: one of them includes the proposition that the railroads run roughly north-south, the other one doesn't. When railroads are under discussion, the fragment that does not have the railroad tracks running north-south is active. This fragment gives an answer to the question that is incompatible with the one the other fragment gives. Thus, we have a case of intuitively inconsistent beliefs. But we can imagine a small variant of the case, where the fragment activated

24 Cf. Lewis 1982; Powers 1978; Stalnaker 1991, *inter alia*.

25 Lewis 1982, p. 436.

26 See Stalnaker 1984, ch. 5 for another version of this suggestion.

for the purposes of discussing railroads in New Jersey is simply undecided as to whether the railroad runs roughly north-south. The moral is that when an agent believes p but fails to believe q, even though p implies q, we should model her belief state by two fragments. According to one of the fragments, p is true; according to the other, neither p nor its negation is. The point of deductive inquiry is (partly) to aggregate one's belief fragments.[27]

This way of thinking about deduction raises a host of difficult questions that are beyond the scope of this paper.[28] But it is a promising strategy for thinking about deductive inquiry that is motivated by a natural way of understanding the phenomena. In what follows, I will assume that it is on the right track. I want to build on this model to give an account of deductive reasoning in mathematics that is compatible with my proposal.

On the view I favor, fragments can differ not only in what worlds they take to be possible, but also in how those worlds are carved up—in other words, in what propositions make up the fragments' hypothesis spaces.[29] So fragments can differ not only in what factual beliefs they include, but also in what mathematical theories they accept—for accepting a mathematical theory is a matter of carving up logical space in a particular space.

Now, one reason fragmentation is attractive to the case of straightforward beliefs is that, while each fragment is perfectly consistent, they are in conflict with each other. An agent who has contradictory beliefs is to some extent defective. Modeling her cognitive state by a fragmented belief system captures a sense in which her beliefs are somehow defective: the fragments are inconsistent with each other. There is a lack of unity in her picture of the world.

---

27 In light of general results in social choice theory and the theory of judgment aggregation, it is safe to conclude that there will be no easy answer to how exactly such aggregation should proceed. (See e.g. List 2008.) But this should not come as a surprise. As Gilbert Harman has often pointed out, after realizing that p follows from q an agent who believed p can either come to believe q or instead abandon her belief in p. It is an open question which option she will take. Cf. Harman 1986 on the distinction between inference and implication.

28 In particular, we need to get clear on how fragments are to be individuated.

29 See Yalcin 2008 on how partition-sensitivity can be used to motivate this model of deductive reasoning.

But consider an agent who accepts two inconsistent mathematical theories. I am suggesting we model her belief state by two fragments: each one would be partitioned by the conceptual resources generated by one of the theories. Again, an agent who accepts inconsistent mathematical theories is in some way defective. Yet it is hard to see what conflict there could be between two fragments that divide the same set of possibilities in different ways. Why not think of the agent as having one belief system consisting of the given set of possibilities and containing each of the propositions available at either fragment?[30]

Perhaps our agent makes use of distinct hypothesis spaces in different contexts. But while this might serve as a motivation for thinking of different partitions of logical space as belonging to different fragments, it is not enough to do justice to the phenomena. When an agent's descriptive beliefs are inconsistent, we are often *forced* to treat her system of beliefs as fragmented. The different fragments are in genuine conflict with each other. Assume we posit fragmentation because our agent uses different hypotheses spaces in different contexts, but suppose both fragments agree on which worlds are possible. Can we nevertheless claim that the two fragments are in conflict with each other?

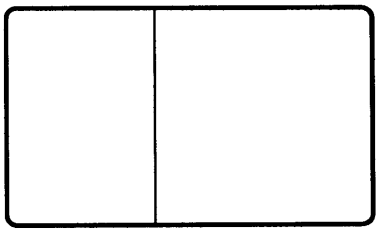
One hypothesis is that the conflict arises out of limitations in an agent's cognitive resources. Our agent may be unable to incorporate the two partitions into an all-purpose one. But there is a deeper reason why different fragments may be in conflict even when they agree on what worlds are possible.

To see that, let me go back to an aspect of the Königsberg bridges example that I set aside in §2.2. Recall that Euler's solution could be split in two steps: first, the introduction of graph-theoretic resources; second, the realization that the answer to (KB) is *no*. The second step is where deduction comes in.

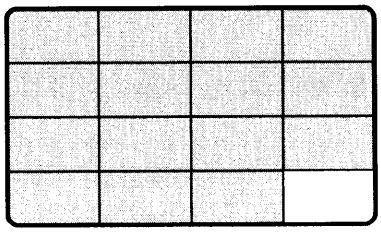
Plausibly, Euler already knew that each landmass in the city of Königsberg was reached by an odd number of bridges. And this, we now know, entails that the answer to (KB) is no. So we have a simple case of failure of deductive closure, one to which we can apply fragmentation.

30 This worry is related, although subtly distinct from, the so-called *negation problem* for metaethical expressivism. See Dreier 2006a; Schroeder 2008b for discussion and references, as well as the discussion in Chapter 1.

We can think of the two fragments as in Figure 2.7. The first is carved up by the two answers to (KB), and contains possibilities corresponding to each answer. The second is carved up by the answers to the question: which landmass is reached by an even number of bridges? Here, we can assume that the only answer compatible with Euler's beliefs is *none*.



(a) Possible configurations of the city, divided by the proposition that the answer to (KB) is 'no'.



(b) Possibilities classified according to which landmasses are reached by an even number of bridges.

Figure 2.7: Two fragments of Euler's cognitive state. The only answer to (KB) compatible with the fragment in (b) is *no*, since according to it the proposition that no landmass is reached by an even number of bridges (the one not grayed out in the figure) is true. The fragment in (a), in contrast, does not settle (KB).

At first, Euler was unable to use his knowledge about how many landmasses are reached by an even number of bridges to answer (KB). This can be represented by a fragmented belief state. In this case, different possibilities are compatible with each fragment. Possibilities in which all landmasses are reached by an even number of bridges are compatible with one of the fragments (the one that is carved up by the answers to (KB)), but incompatible with the other.

But go back to a point *before* Euler realized that each landmass is reached by an odd number of bridges. We can assume, to make things simpler, that his non-mathematical beliefs were deductively closed. He did not know what the answer to (KB)) was, but he also didn't know how many bridges reached each landmass in the city. Nevertheless, I submit, we should model his cognitive system as fragmented. For he was disposed to have a fragmented belief state: he was disposed to form the belief that there was an odd number of bridges reaching each of the landmasses *without* forming the belief that the answer to (KB)) was *no*.

And this is because he was unable to use evidence that could settle the questions carving up one fragment—is each landmass reached by an odd number of bridges?—in order to answer the questions carving up the other—viz., (KB)). He was unable to see how hypotheses from the two fragments relate to one another.

Deductive reasoning can eliminate inconsistencies in one's descriptive beliefs. But it can also improve one's *information transfer abilities*. An agent whose descriptive beliefs are in conflict with each other hasn't transferred information from one fragment to another. An agent whose mathematical views are inconsistent is *disposed* to be in that situation. This is because she hasn't acquired the ability to use evidence settling one question to answer another, logically related one.

Note that once we model things this way, we can see how the discovery of graph-theory could have helped with determining that the answer to (KB) is *no*. Suppose you could partition logical space in such a way that it was easy to tell (i) which cell of that partition the actual world belongs to and (ii) what the answer to (KB) was, given the answer to (i). The coarse partition given by (KB) itself makes (ii) trivial, but is of no help with (i). The fine partition given by detailed descriptions of the city might make (i) easy, but not (ii). Euler's accomplishment in introducing the theory of graphs was to provide a partition meeting both (i) and (ii).

Indeed, we can think of this partition as mediating the transition from the one fragment in Figure 2.7 to the other. This is the cognitive accomplishment involved in the proof of Euler's theorem: establishing a bridge from the proposition that an odd number of bridges reaches each landmass in Königsberg, through Euler's proposition, to the proposition that the answer to (KB) is *no*. The transition from the former to Euler's proposition, and that from Euler's proposition to the latter can each be seen as simpler, more immediate ones (cf. Figure 2.8). Euler's lack of logical omniscience was manifested in his inability, before proving the theorem, to transfer information from the fragment triggered by the question 'how many landmasses are reached by an odd number of bridges?' to the fragment triggered by (KB).

The interaction between fragmentation and my account of mathematical thought can thus be used to illuminate the way in which mathematical theorizing can increase our computational resources. But it can also be used to sketch an account of the role of deductive reasoning

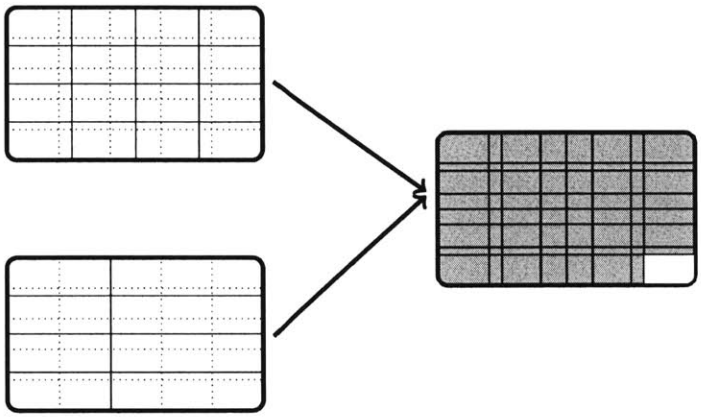


Figure 2.8: Refinements can be thought of as helping to calibrate two fragments. By arriving at a common refinement of two fragments, it is easier to see how cells in each fragment relate to cells in the other.

in mathematical thought.

2.6 OUTSTANDING ISSUES

Many issues remain outstanding. Here are two of the most pressing ones.

First, is my proposal compatible with a plausible semantics for mathematical language? To give a full account of mathematical practice we certainly need to give a compositional semantics for the relevant fragments of our language. On my view this needn't be the first step: a different starting point can give us a more illuminating theory. Formal semantics is often seen as a neutral ground on which disputes about the nature of mathematical practice should take place. But I see no reason why semantics should be the royal road to understanding our mathematical practice. The complexity of this practice goes beyond anything that can be explained by giving a semantics for a fragment of natural language. And focusing on the details of a compositional semantics might make us lose perspective.

If I am right, the goal of mathematical practice is to arrive at new ways of carving up logical space. But we engage in inquiry as a community, and we need to agree on how to carve up logical space in order to communicate with each other. It would be surprising if we did not have

some way of fostering such coordination. A language is well-suited to do this: it is a device for expressing our mental states and trying to arrive at some common state. The difficulty is to give a detailed story of how *our* mathematical language could be understood as playing that role. I suspect that recent work done on nonfactualism—in particular (Gibbard, 1990, 2003)—will be helpful for sketching a semantics for a fragment of our mathematical language. However, this is a task for some other time.[31]

Second, can my proposal be generalized to deal with mathematical theories that are more abstract—theories involving large cardinal axioms, say, or theories that seem non-applicable to the natural sciences? I don't have an answer to this question, yet. But here is one line of thought worth exploring. In the same way that accepting 'lower-level' mathematical theories can be seen as adopting ways of making new distinctions among ordinary, descriptive propositions, accepting 'higher-level' mathematical theories can be seen as making new distinctions among lower-level theories. We can motivate this strategy by noting that more abstract branches of mathematics often arise out of reflection on more 'concrete' mathematical theories. This gives us a hierarchical picture, on which the kind of distinctions we make at one level will be constrained by the theories we accept at higher levels.

2.7 CONCLUSION

Most attempts at making sense of our discursive practices proceed in full generality. They ask what asserting a declarative sentence is, or what judging that so-and-so amounts to. A less ambitious strategy—favored by metaethical expressivists—is to leave open the possibility of treating different domains of discourse differently. The strategy is to ask what it is to make a *moral* judgment, or a *mathematical* judgment, not simply what it is to judge that so-and-so. This is the strategy I have adopted. It would be nice if we could say something general about all sorts of discourses. Reflection on our discursive practices suggests that they

31 It is not hard to sketch such a semantics based on the notion of *scorekeeping*, much along the lines of Lewis 1979b and Stalnaker 1973. The main observation—which I develop in the appendix—is that we can easily represent the evolution of the part of the score corresponding to the partition presupposed by conversational participants as proceeding by 'elimination' of alternatives.

have much in common. But this is not a nonnegotiable constraint on the project.

In this paper, I defended a novel account of mathematical practice. On this account, to discover a new mathematical theory (or structure) is not to acquire a new belief. Rather, it is to change the granularity of one's working picture of logical space—in other words, to change one's working hypothesis space. Discovering a new mathematical theory involves acquiring the ability to *see* any possible physical configuration as a potential instance of the theory.

The picture that emerges from my proposal is a form of nonfactualism. But it is one that can account for the ways in which our concern for the truth imposes substantial constraints on our mathematical theorizing. For how best to structure our inquiry into the physical world will depend on what our epistemic goals are. This opens the door to an account of the rationality of our mathematical practice that is compatible with a plausible picture of our cognitive lives.

2.A APPENDIX: CONTENT AND SEMANTICS

Distinguish two questions:

(i) What is the functional role of a particular type of mental state?

(ii) What is the most perspicuous way of representing those mental states in a theory of the mind?

I have said something in answer to the first question. But what I have said does very little to constrain how we should answer the second one. Let me explain.

When I want a break, I am in a particular mental state. We might find it convenient to explain what that state *is* by appealing to the functional role it plays. We might say that it is a state that tends to bring it about that I look for an excuse to stop working, or that I get up and walk to the water cooler. This is no doubt a very partial characterization of what state I am in, but we can give a theory of *wants* along these lines.

Yet when deciding how to represent *wants* in our theory of the mind, we might want to use *propositions*. We might want to represent my wanting a break as my being in a particular relation to the proposition that I take a break. But this is not *forced* upon us by our account of what *wants* are. Rather, it is something we do in order to streamline our theory of the mind. For in assigning propositions as objects of my wants, as we do with beliefs, we can make useful generalizations about the way in which our wants and beliefs interact with one another, by looking at the logical relations between the propositions we assign to them.[32]

Sometimes we have additional pressure to represent attitudes as propositions. Consider attitudes like *expectations*. We not only expect rain, we also expect *that* it will rain. It seems natural to represent the latter state as involving a relation to a proposition (the proposition that it will rain), and it would be odd not to do the same with the former.

On my view, mathematical beliefs do not involve eliminating possible worlds. Rather, they involve taking on the conceptual resources

32 This point is nicely made by David Lewis (1979a, §1): "Our attitudes fit into a causal network. In combination, they cause much of our behavior (...). In attempting to systematize what we know about the causal roles of attitudes, we find it necessary to refer to the logical relations among the objects of the attitudes. Those relations will be hard to describe if the assigned objects are miscellaneous."

that come with the relevant mathematical theory. But our mathematical beliefs relate to one another much in the same way that our ordinary, descriptive beliefs relate to one another. And our best way of thinking about the way in which our ordinary, descriptive beliefs relate to one another involves representing them as relations to propositions. It would be nice, therefore, if we could also represent mathematical beliefs using proposition-like objects. For then we would be able to transfer all we know about our theorizing about ordinary, descriptive beliefs to our account of mathematical thought.

Fortunately, a simple trick will let us do so.

2.A.1 *Content*

Assume for a moment that we can associate with each mathematical structure a partition of the space of possibilities—i.e., a collection of propositions that are mutually exclusive and jointly exhaustive. This will correspond to the smallest set X of propositions such that any proposition that can be entertained given the conceptual resources the theory provides is a Boolean combination of elements of X. For example, the proposition that describes the graph-theoretic structure of the city of Königsberg is one that can be isolated using the graph in Figure 2.2—we can, as it were, point to that graph and say that the structure of the bridges is like *that*. I will speak of a proposition as being 'generated by a structure' whenever it is one of those built out of the partition associated with that structure.

Now, we can characterize the state of accepting a mathematical theory with a set of mathematical structures: those structures such that they each generate the (structural) propositions that make up our agent's picture of logical space. As we will see, we can think of those structures as being precisely what we normally think of as the *models* of the theory (in the model-theoretic sense).

If our agent comes to accept a claim that is independent of the theory she accepts, she will have narrowed down the set of structures that characterize her mental state—much like in forming a new descriptive belief an agent narrows down the set of worlds she takes to be possible. Which structures will be ruled out? Those that don't allow for the conceptual resources that the new theory generates.

A picture might help. Take a look at Figure 2.9, and think of i, j and

k as standing for different mathematical structures. Each layer in the picture corresponds to a different way of partitioning logical space—$Q(i)$ corresponds to the partition generated by i, and so on. An agent that has not settled on which of $Q(i)$, $Q(j)$ or $Q(k)$ to take on as her working hypothesis space can be modeled by the set containing i, j and k. In particular, such a set would model our agent having adopted p, as part of her hypothesis space, but not yet q nor r. If she went on to include r as part of her picture of logical space, she could be modeled as having ruled out i, since it does not generate r (in other words, r is not a member of $Q(i)$).

Such a change could be represented as the transition from a partition that includes all and only those propositions in each of $Q(i)$, $Q(j)$ and $Q(k)$—which we can denote by $Q(\{i,j,k\})$—to one that includes all and only those propositions in each of $Q(j)$ and $Q(k)$—which we denote by $Q(\{j,k\})$. Thus, such a change could be seen as an expansion in our agent's conceptual resources that can be represented by the elimination of one mathematical structure. You can look at the contrast in Figure 2.10 for illustration.

Note the analogy with our representation of changes in an agent's ordinary, descriptive beliefs. When we model an agent as having ruled out a possibility, it is because she comes to believe *more* propositions. Similarly, we can model our agent as ruling out a mathematical structure when she comes to take on *more* propositions as being part of her hypothesis space.

If we want to be able to assign proposition-like objects not only to 'pure' mathematical beliefs, but to 'mixed' ones as well, we can simply model an agent's cognitive state as a set of *pairs* $\langle w, i\rangle$,[33] where w is a possibility and i is a mathematical structure used to encode the conceptual resources we ascribe to an agent. Now suppose we represent an agent's cognitive state by a set of pairs H. We can read off from the set H which worlds are compatible with our agent's beliefs *and* which possible world propositions make up her hypothesis space. A possible world proposition A will be in our agent's hypothesis space iff it is in $Q(i)$ whenever i figures in a pair that is in H. A world w will be compatible with her beliefs just in case w figures in some pair that is in H.

To complete this sketch, we need to discharge the initial assumption

33 Cf. Gibbard (1990, 2003).

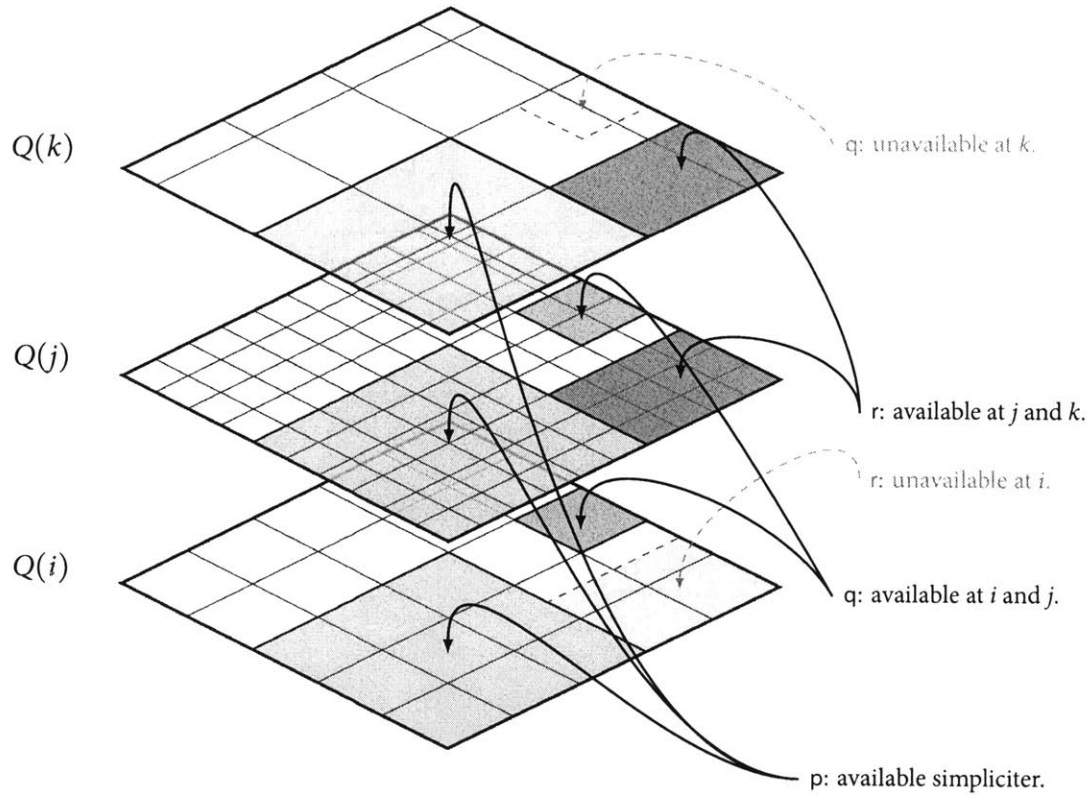
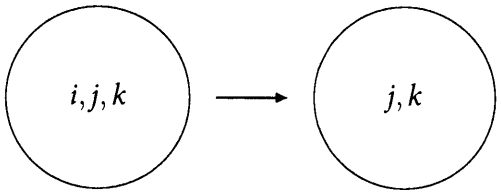
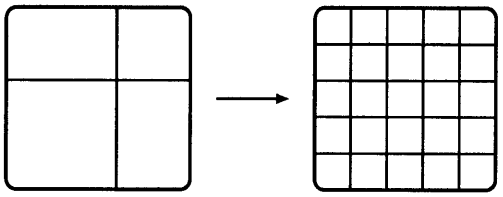


Figure 2.9: Alternative conceptual resources: This represents an agent that has p among her working hypothesis space, but has not yet taken in either q or r. We can model her coming to take on q as the result of her 'ruling out' *k*.



(a) Ruling out points.



(b) Refining logical space. The transition from $\{i, j, k\}$ to $\{j, k\}$ is used to stand for a transition from $Q(\{i, j, k\})$ to $Q(\{j, k\})$.

Figure 2.10: Two ways of modeling one change in our agent's cognitive state

that there is a good assignment of sets of propositions to each mathematical structure. Here is how. For any mathematical structure and any physical system, we can ask whether there is an isomorphism from the mathematical structure and the physical system.[34] Any proposition (or the negation of a proposition) of the form 'S is a physical system isomorphic to M' can be thought of as generated by M, as will any finite Boolean combination of such propositions. In other words, the set of propositions assigned to a mathematical structure M will be the largest collection of propositions A such that for any two worlds w and w' in A, if a physical system in w is isomorphic to M, so is its counterpart in w'. To take an example, recall our graph in Figure 2.2: it can be used to generate the (true) proposition that the bridges of the city of Königsberg are isomorphic to it, the (false) proposition that the bridges of the city are not isomorphic to it, the (false) proposition that the bridges of Paris are isomorphic to it, the proposition that the faucets and pipes of a given house are isomorphic to it, and so on.[35]

34 We can also ask whether there is a partial isomorphism, i.e. an isomorphism from a substructure of the given structure to the given physical system.

35 Alternatively: it can be used to generate the proposition consisting of those worlds in

We can thus use sets of subsets of I to represent what mathematical theories our agent accepts. By assigning to an agent a set $H \subset I$ as the 'content' of her mathematical beliefs, we attribute to them the conceptual resources that correspond to the intersection of those X_i such that $i \in H$. This means that we can use proposition-like objects to represent our agent's mathematical beliefs *even if* to have a belief represented by a set of points H is not a matter of taking a stance on what the world is like, but rather a matter of deploying certain conceptual resources in theorizing.

2.A.2 *Semantics*

A further advantage of this representation is that it allows us to give an adequate semantics for the language of mathematics.

To give a compositional semantics for a given language is to assigning in a recursive fashion an abstract object to any well-formed sentence in the language. On my view, the main constraint on this project is that these abstract objects can be used to model the effect of an utterance of that sentence in a conversation. So in order to give a semantics for the language of mathematics, we need to answer three questions.

(Q1) What is the effect of an utterance of any such sentence in a conversation?

(Q2) What objects can be used to characterize the effects of utterances in a conversation?

(Q3) How can we assign the relevant objects to well-formed sentences of the language in a recursive fashion?

I will answer each question in turn. Before I do that, I want to sketch a well-known abstract model of conversation. This will set the stage for what follows.[36]

Think of a conversation as an activity whose purpose is to induce changes in the mental states of the participants. Each stage in the conversation can be characterized by the *conversational score*, which represents the states of mind that speakers take the participants to be in. The effect of an utterance can be captured by a transition rule: a rule

which the bridges of Königsberg are isomorphic to the graph in Figure 2.2, etc.

36 Cf. Lewis 1979b; Stalnaker 1973.

that tells us how the conversational score should change if that utterance is accepted. The simplest example of this model is one in which the conversational score is just a set of possible worlds. This set contains those worlds that are taken to be live possibilities for the purpose of the conversation: they are the worlds that represent what speakers take each other to presuppose. To any sentence, we can assign a set of worlds. Uttering any sentence will eliminate, from the conversational score, those worlds not in the set of worlds assigned to that sentence.

The score can be more complex if there are different mental states we want to keep track of. Suppose we have a language that can induce changes in what speakers take each other to believe, and independently induce changes in what speakers take each other to accept as a standard of precision (i.e. whatever determines whether an utterance of a sentence like 'France is hexagonal' is appropriate). We will then want to have a score containing two elements: a set of worlds, and some abstract object that can encode information about what standards of precision are relevant for the conversation.

We can now go back to our three questions. First: how does uttering a sentence with mathematical vocabulary affect a conversation? On my view, to accept a mathematical theory is to adopt a particular way of structuring logical space. The point of uttering a sentence with mathematical vocabulary is in part to induce changes in what way of structuring logical space to accept for the purposes of the conversation—in what the hypothesis space of the conversation is.

What objects can be used to characterize the effect of these utterances? For any mathematical language $\mathcal{L}$, we can use subsets of the collection of $\mathcal{L}$-structures to characterize the hypothesis space of the conversation, using the construction in §2.A.1. Since the dynamics of refining logical space can be represented by 'narrowing down' that set, we can simply assign a set of mathematical structures to each sentence. Uttering that sentence will rule out the mathematical structures that are not in the set associated with the given sentence.

Finally, how can we assign these objects to our sentences in a systematic way? We can do so just as a descriptivist would. We can proceed by assigning to each sentence in a mathematical language $\mathcal{L}$ the set of $\mathcal{L}$-structures that are models of the sentence. This assignment can be done much in the same way that we do semantics for any first-order language.

It is slightly trickier to give a semantics for *mixed* sentences. Here is one promising way to do so. Start by having our conversational score contain a set of worlds and a set of mathematical structures. Plausibly we can assign a set of *pairs* of the form $\langle w, i \rangle$ to each such sentence: the set of $\langle w, i \rangle$ such that there is a (partial) isomorphism between the physical system—in w—that the sentence is about, and the structure i. Again, we can piggy-back on a relatively straightforward assignment of a set of pairs of a world and a mathematical structure to each mixed sentence. We can use these sets in order to represent the effect that any mixed sentence will have on the conversational score, roughly along the lines suggested by the construction in §2.A.1.

On my view, an assignment of semantic values to sentences places fewer constraints on our theory of the relevant mental states than is usually supposed. All it tells us is that the algebra one uses to provide the semantics for a language is isomorphic to the algebra one uses to characterize the relevant mental states, and the way these evolve during a conversation. But—and this is the crucial point—-there is more than one way of understanding the role of the algebra that we use to characterize our mental states. For example, on a conventional interpretation, for a state s to be weaker (in the sense of the algebra) than state s is for the *truth-conditional content* of s' to entail the *truth-conditional content* of s. On my interpretation, in contrast, for s to be weaker than s' is for s' to deploy richer conceptual resources than s. The upshot is we can give a rather non-revisionary semantics for the language of mathematics. Crucially, we can do so without assuming that to accept a mathematical theory is to take a stance on what the world is like.

GOOD QUESTIONS

We care about the truth: we want to believe what is true, and avoid believing what is false. But not all truths are created equal. Having a true botanical theory is more valuable than having true beliefs about the number of plants in North Dakota. To some extent this is fixed by our practical interests. We may want to keep our plants looking healthy, and doing botany is more likely to help us do that than counting blades of grass. (Of course, if you find yourself out of luck, your main goal in life might be to keep tally of blades of grass.) But setting our practical interests aside, there is something more valuable, *epistemically*, about our botanical beliefs than about those we get out of counting blades of grass.

Now, you might worry that there is not much content to the notion of *purely* epistemic value. Perhaps any epistemic dimension of evaluation will be somewhat entangled with pragmatic considerations.[1] But we can surely evaluate our beliefs so as to minimize the interference of pragmatic considerations. We can set aside particular idiosyncrasies of our judgments of practical value, and focus instead on how some beliefs more than others help us make sense of the world around us.

That, at least, is the intuition driving this paper. I think it is a powerful intuition—one that has yet to be cashed out without relying too heavily on metaphors. The task of the paper will be to do just that. More specifically, I want to offer way of evaluating different courses of inquiry—different research agendas, as it were—from a purely epistemic perspective. There will be some additional benefits from looking at things the way I suggest. But I will not get to that until the end.

A word about methodology is in order. I will speak interchangeably of 'beliefs' and 'credences', or 'degrees of confidence'. I realize there are some difficult questions about how those two notions relate to one another, if at all.[2] For our purposes, however, we can set them aside. Most

1 For more on skepticism about the notion of a purely epistemic notion of value, see Gibbard 2008, as well as Arntzenius 2008.

2 For discussion, see e.g. Sturgeon 2008.

of what I will say can be recast purely in terms of all-out beliefs—only towards the end will the notion of a credence function play a major role. For now, those details do not matter.

3.1 EVALUATING QUESTIONS

We want to evaluate truths. But better to first evaluate lines of inquiry, or research agendas. To think that p would be good to believe, if true, is at least to think it is worth engaging in finding out *whether* p—at least if we set aside, as I will, the cost of inquiry, and the likelihood that you will determine whether p. To the extent that you think it is worth trying to answer the question of whether p, you will think that at least one of its answers would be good to believe, if true.

We cannot just identify the value of p with the value of answering the question whether p. This would have as a consequence that you will think of every proposition as valuable as its negation. And while a sophisticated scientific theory might be very valuable, epistemically, its negation may not be worth much. Keeping this in mind, however, let us start by evaluating lines of inquiry. Connecting issues about the value of truths to issues about the value of *actions* gives us enough tractability, I believe, so as to outweigh any reservations you might have. And since what we can do is decide whether to engage in finding out whether p, this is a good place to start.

Now, you can think of a course of inquiry as a collection of questions. Indeed, you can think of it as a single question: what is the answer to each of these questions? So any way of evaluating questions will thus correspond to a way of evaluating courses of inquiry.

We can easily devise a framework for evaluating questions using fairly simple decision-theoretic tools. We need only assume that we can identify the value of a question with the expected value of *learning* the answer to that question. For whenever you are facing a choice among a set of options, you can evaluate questions according to how likely, and to what extent, learning its true answer will help you make the right choice.

Let's cash this out a bit more explicitly. Two coins will be tossed. You are told that the first coin is fair. The second one is biased: there is a 70% chance it will land heads. Consequently, you assign credence .5 to the first coin landing heads, and .7 to the second one landing heads. You are

then asked to predict a particular outcome: you will be rewarded only if you predict the actual outcome. The reward will depend on what the prediction is, according to this table:

HH: $0	HT: $5
TH: $10	TT: $15

After computing the expected utility of each possible action, you realize that TH is the action that maximizes expected utility.[3] Before you state your choice, however, you are told you can ask the Oracle one of two questions:

(Q1) Did the first coin land heads?

(Q2) Did the second coin land heads?

Clearly, if you have nothing to lose, you should ask one.[4] The question is: which one?

To answer this, we need to consider two different issues. First, all things being equal, we prefer to ask one question over another if we are less opinionated about the answer to the first question. If we have good evidence that the answer to Q is a, but no evidence pointing to what the right answer to Q' is, we have a *pro tanto* reason for asking Q' rather than Q. At the same time, if we expect that having an answer to one question will have little impact on our choice—perhaps we would choose the same action no matter what the answer to that question is—we may have reason to ask another question instead of that one. We need a way of arbitrating between these potentially conflicting considerations. Following I. J. Good (1967), we can do so in the following way.

First, for each possible outcome of the coin tosses we estimate the net gain in utility you expect from acting after learning that outcome. This is the difference between the expected value, relative to the result of updating your credence function (your *posterior*), of (i) the action that

3 Your credence assignment is as follows: $C(\text{HH}) = C(\text{TH}) = .35$, $C(\text{HT}) = C(\text{TT}) = .15$, so that the expected utility of TH is $3.5, whereas that of TT is $2.25.

4 We know from a result by I. J. Good that for any Q (and any decision problem) the value of asking Q is never negative, *so long as asking Q is cost-free.* See Good 1967. Good attributes the result to Raiffa and Schlaifer 1961. For a general discussion of Good's theorem, as well as a generalization of the result when different updating rules are allowed, see Skyrms 1990.

maximizes expected value relative to the posterior, and (ii) the action that maximized expected value relative to your prior.[5] We can then identify the value of a question with the weighted average of the values of its possible answers.

Return to the coin example. Relative to your prior credence function, TH was the action that maximized expected utility. But if you learned that the first coin landed heads (henceforth, 'H1') you would no longer pick TH. For assuming you update your credence function by conditionalizing on your evidence, that would be a sure loss. The sensible thing to do if you learned H1 would be to pick HT, since the expected utility (relative to your new credence function) of each other option is \$0. Now, the expected value of HT relative to the result of updating your credence function with the information at hand is \$1.5. Since upon acquiring that information the expected value of TH would be \$0, the net gain in utility from learning that the first coin landed heads is $V(\text{H1}) = \$1.5 - \$0 = \$1.5$.

Similarly, we can compute the expected gain in utility from learning that the first coin landed tails (i.e. T1): it is the expected value of TH relative to the posterior, minus the expected value of TH also relative to the posterior. In short, learning T1 would not affect your choice, so that $V(\text{T1}) = 0$.

We can then assign as value to Q1 the weighted average of the values of its answers, so that $V(\text{Q1}) = \$.75$.[6] It is easy to verify that $V(\text{H2}) = \$\,0$, and $V(\text{T2}) = \$\,7.5$. This allows us to assign a value to Q2 as the weighted average of the value of H2 and T2, i.e. $V(\text{Q2}) = \$2.25$.[7] The upshot is that the value of Q2 is higher than that of Q1, so that Good's strategy recommends you ask Q2, as we would expect.

I want to use this strategy to spell out a way of evaluating questions from a purely epistemic perspective. But I first need to find the right decision problem.

---

5 As it turns out, one could also assign value to a proposition p by looking at the difference between the posterior expected value of the action that maximizes expected value relative to the result of conditionalizing on p and the expected value of the action that maximizes expected value relative to your prior. The value of p will of course be different if we do things this way, but the resulting $V(Q)$ will be the same. See van Rooy 2004, p. 397.

6 Since $C(\text{H1}) \times V(\text{H1}) + C(\text{T1}) \times V(\text{T1}) = .5 \times \$1.5 + .5 \times \$0$.

7 Since $C(\text{H2}) \times V(\text{H2}) + C(\text{T2}) \times V(\text{T2}) = .7 \times \$0 + .3 \times \$7.5$.

3.2 EPISTEMIC DECISION PROBLEMS

I have implicitly appealed to a very minimal form of *expected utility theory*. On this framework, different alternatives are evaluated relative to a credence function and a *utility function*—an assignment of numerical values to each alternative in any possible world. A particular alternative is *better* than another if it has a higher expected utility—again, relative to a credence and utility functions.

The canonical application of this framework is to give an account of rational decision theory—to solve *decision problems*. Typically, a decision problem is just a set of alternative courses of actions. We evaluate these relative to a given utility function and a credence function. On one view, an agent's actions are *rational* just in case they have the highest expected utility (among a relevant set of alternatives) relative to her own credence function *and* her own utility function.

But we can apply this conception to rationality to other situations. Whenever we have a range of options and an assignment of utility to each option relative to each possible state of the world, we can apply expected utility theory to evaluate each of the relevant options. In particular, we can think of decision problems where the alternatives are possible epistemic states one could be in. So long as we have a credence function (defined over possible states of the world) and a utility function defined over the relevant alternatives (relative to a possible state of the world), we can use expected utility to compare different possible epistemic states.

This way of evaluating alternatives is relativistic in an important sense: it only makes sense to ask whether a given option is better than another relative to a particular credence function and utility function.[8] But this does not prevent us from using it to capture thicker notions of value: we only need to make more restrictions on what is to count as an admissible utility function (ditto for credence functions).

In particular, we can use it to capture a notion of *epistemic value*. Given a credence function, an 'epistemic utility function', and a set of epistemic states, we say that an epistemic state is better than another, *epistemically*, iff it has higher expected utility relative to that credence and utility functions. But for this to be of any help, we need to spec-

8 Cf. Stalnaker 2002, p. 158.

ify what a utility function must be like if it is to count as an *epistemic* utility function—a utility function that corresponds to an epistemic dimension of evaluation.

To give you a sense of the kind of evaluation I'm after, consider the following case.

> An Oracle tells you the truth-value of every proposition. She then tells you that you will be put to sleep and your memory will be erased. Fortunately, you can now pick which credence function you will wake up with.

If you could pick *any* credence function, I suppose you would know what to choose: the one that assigns 1 to all and only the true propositions. But here is the catch: you cannot pick any credence function. You will be given a choice among a small set of credence functions which does not include the one you currently have. If you set aside your practical interests for a moment, how will you choose?

I suspect you would still have a rough idea of how you would choose. You would be able to compare at least *some* epistemic states with each other, in a way that would correspond to an epistemic dimension of evaluation. Intuitively, a utility function will count as an *epistemic* utility function just in case it corresponds to the way a fully informed agent would rank epistemic states from an epistemic perspective.

To be sure, this cannot be a *definition* of an epistemic utility function, unless we have a clear enough notion of what an epistemic dimension of evaluation is. But it is a useful heuristic, one that can help motivate conditions that must plausibly be met by anything that could count as an epistemic utility function.

3.2.1 *Truth-directedness*

Here is the first thing we can say about what an epistemic utility function must be like. Suppose you are comparing two credence functions C and C', defined over a body of propositions $\mathcal{B}$. If you know which propositions in $\mathcal{B}$ are true, then you will probably think that C is better than C', epistemically, if for each $p \in \mathcal{B}$, $C(p)$ is closer to p's truth-value than $C'(p)$ is.

We can extract a minimal constraint on epistemic utility functions from this. Recall that utility functions are assignments of numerical

values to each pair consisting of an alternative—from the relevant decision problem—and a possible world. If alternatives are just credence functions, our utility functions in question assign numerical values to pairs of the form (C, w), where C is a probability function defined over a fixed $\mathcal{B}$, and w is a possible world. Say that C is uniformly closer to the truth than C', relative to w, if for each proposition over which p is defined C assigns a value to it at least as close to its truth value in world w than C' does, and it gets closer than C' for at least one such proposition. The constraint that must be satisfied by any epistemic utility function can now be stated as:[9]

> TRUTH-DIRECTEDNESS: If C is uniformly closer to the truth than C', then $u(C, w) > u(C', w)$,

To illustrate, suppose you are only interested in one question: whether Secretariat won the Kentucky Derby in 1973. You are told by a reliable source that he did. If you are given the choice between waking up with a credence function that assigns .6 to the proposition that Secretariat won that race, and waking up with a credence function that assigns .9 to the proposition that Secretariat won that race, you would presumably choose the latter. After all, it gets closer to the truth of the proposition you are interested in, and you have nothing to lose. TRUTH-DIRECTEDNESS constrains epistemic utility functions to agree with your judgment: from an epistemic point of view, in a world in which Secretariat did win the race in 1973, it is better to have a credence function that assigns .9 to that proposition than one that assigns .6 to it.

Note that TRUTH-DIRECTEDNESS only tells you to rank C above C', relative to a world w, if C is closer than C' to the truth-value, at w, of *every* proposition. But suppose you were given two credence functions C and C' to choose from when you wake up. You now know the truth-value of p and q. If C is closer to the truth about p but C' is closer to the truth about q, TRUTH-DIRECTEDNESS will tell you nothing about how to chose, even if C is much closer to the truth about p than C' is about the truth about q.

To evaluate different lines of inquiry we need a principled way of going beyond TRUTH-DIRECTEDNESS in precisely those contexts. Do-

9 Cf. Joyce 2009.

ing that will be the object of §3.4.[10] Before moving on, however, I want to consider an additional constraint on epistemic utility functions, one that is almost as uncontroversial as TRUTH-DIRECTEDNESS.[11]

3.2.2 *Propriety*

Suppose you are rational in having credence function C. Further suppose you are evaluating alternative credence functions relative to your own credence function and an epistemic utility function u. Could it be that you take some *other* credence function to be better, epistemically, than your own?

The question is not whether you could reasonably think that some of your beliefs could be false. The question is whether you could reasonably think: it would be better, from a purely epistemic perspective, to have C' as my credence function. If you are rational, it seems, you could not. So if you have rationally arrived at your credence function C, then the expected epistemic value of C' should not be greater than the expected epistemic value of C.

If we further assume that any credence function could be rationally held by some agent at some point, then we will be tempted to endorse the following condition on epistemic utility functions:

> PROPRIETY: For any credence function C, the expected value of C relative to u and C must be greater than or equal to the expected value of C' relative to u and C.

PROPRIETY ensures that a probability function will always evaluate itself, relative to u, as having maximum expected value.[12] In fact, one

10 Thus, I will be suggesting that epistemic utility functions need not be *normal*, in the sense defined in Joyce 2009. Joyce considers normality as a constraint on *accuracy* measures, and takes no stand on whether to accept it. And I suspect he would agree that accuracy is not all there is to epistemic utility.

11 Other constraints have been proposed. For example, Joyce 1998 discusses a constraint he calls LOCALITY, which makes $u(C, w)$ be sensitive only to the value of $C(p_w)$, where p_w is the strongest $p \in \mathcal{B}$ such that $p(w) = 1$.

12 By itself, TRUTH-DIRECTEDNESS is not enough to guarantee PROPRIETY. For instance, take the following value function defined over all credence functions whose domain is a given finite set X of size N:

$$\varphi(C, w) = (-1)\frac{1}{N}\sum_{p \in X} |C(p) - p(w))|.$$

might want to impose a slightly stronger constraint on epistemic utility functions, viz. that the expected value of C relative to C and u must be *higher* than that of any other C'. But we need not argue over that. What matters is that we have some idea of what an epistemic utility function must be like. It is time to put this to use to start answering our initial question.

3.3 THE EPISTEMIC VALUE OF QUESTIONS

I set out to find a decision problem that could allow us to evaluate questions from an epistemic perspective. So far, all I have is a sketch of such a decision problem: a decision problem where options are assessed relative to an epistemic utility function. How could questions be evaluated given such a problem?

Again, we set the value of a question as the weighted average of the value of (learning) its answers. The value of each answer is obtained as follows. First, let a be the alternative that maximizes expected value relative to your current credence function. Now let a' be the alternative that maximizes expected value relative to the result of updating your credence function with the proposition p. The value of (learning) p is the difference in expected value, relative to the result of updating your credence function with p, between a and a'.

I have said very little about what epistemic utility functions are. But assuming that all epistemic utility functions are *proper*, we can at least say this much: the alternative that maximizes expected value relative to your prior C is C. The alternative that maximizes expected value relative to the result of updating your prior with p is $C_p = C(\cdot \mid p)$.[13] So the value you will assign to (learning) p is the difference in expected value, relative to C_p, between C and C_p. And the value you will assign to a question Q is the weighted average of the value you will assign to (learning) each of its answers.

We can simplify things even further, if what we are after is a way of *comparing* different questions. For the expected value of Q will be higher than the expected value of Q' relative to a proper scoring rule

---

Clearly, φ satisfies TRUTH-DIRECTEDNESS. But for most credence functions C, φ will sanction moving to the extremes. See Gibbard 2008 for discussion.

13 Cf. the discussion in Greaves and Wallace 2006 of why conditionalizing maximizes expected utility theory, assuming that all epistemic utility functions are proper.

u and a credence function C if and only if the weighted average of the expected values of C_p relative to u and C_p, for p an answer to Q, is higher than the expected value of C_r relative to u and C_r, where r is an answer to Q'.

But just assuming that an epistemic utility function is proper and truth-directed is not enough to get an answer to our starting question. To see that, suppose you only assign credence to two independent propositions, p and q, as well as their Boolean combinations. Suppose we have an epistemic utility function that is truth-directed and proper, such as[14]

$$u_B(C, w) = -1/2((C(p) - p(w))^2 + (C(q) - q(w))^2),$$

where we identify a proposition p (thought of as a set) with its characteristic function. You are trying to determine which of $?p$ (whether p) and $?q$ (whether q) to ask. If you assign .5 credence to each of p and q, then the expected value of $?p$ will be exactly that of $?q$.[15] So if all we know of epistemic utility functions is that they satisfy TRUTH-DIRECTEDNESS and PROPRIETY, this way of cashing out a notion of epistemic value will not allow for non-trivial comparisons between different lines of inquiry.

The reason our utility function u_B fails to distinguish between $?p$ and $?q$ in this particular situation is that it is insensitive to the *content* of the propositions being assessed. And this is precisely what we want to avoid if we want to compare different lines of inquiry from an epistemic perspective.

Now, we could assign different weights to the terms in the sum above. We could, say, define a scoring rule that would differ from u_B only in that rather than calculating something like average distance from the truth about p and the truth about q, we take a *weighted* average distance, say by multiplying $(D(p) - p(w))^2$ by some factor greater than 1. This would take deviations from the truth about p to be worse than deviations from the truth about q. But we need a good reason for favoring one proposition over the other, and we need to do so in a way that

14 This is the widely studied *Brier score*—cf. Brier 1950.

15 To see that, note that C_p and C_q are perfectly symmetric, so that the expected value of C_p relative to C_p is equal to that of C_q relative to C_q, and the same goes for $C_{\neg p}$ and $C_{\neg q}$.

corresponds to an epistemic benefit to be accrued from being closer to the truth about p rather than closer to the truth about q. In the next section, I want to propose a way of doing just that.

3.4 COUNTERFACTUAL RESILIENCE AND EXPLANATION

You flip a coin ten times in a row. To your surprise, it lands tails nearly every single time. Here is a possible of explanation of what happened:

BIAS: The coin is heavily biased toward tails.

Another possible explanation would consist of a specification of each of the starting positions of the coin and your hand, together with a specification of the force with which you flipped it and the wind conditions which, together with the laws of physics, make it extremely likely that the coin landed tails nearly every time. Call this explanation INITIAL, and suppose it is incompatible with BIAS, perhaps because we can derive from the facts cited in INITIAL that the coin is fair.[16]

To some extent, the first explanation is more satisfying than the second. This is not because it is more or less likely: it may be highly unlikely that the coin you got from the bank was heavily biased towards tails. Rather, it is because *if true* it would be more satisfying as an explanation than the second one would be, if *it* happened to be true.

3.4.1 *Some platitutdes*

Why would an explanation in terms of BIAS be more satisfying that one in terms of INITIAL? It is not because BIAS makes the explanandum more probable. We would prefer BIAS even if we modified INITIAL so that it entailed the truth of the explanandum, and therefore raised its probability to one. Rather, it is because BIAS has some familiar explanatory virtues that INITIAL lacks.

For example, BIAS is simpler than INITIAL. We want our theories to be simple—we want them to involve no more detail than it is necessary—partly because theorizing has cognitive costs, and we rather not

16 This example is based on a slightly different example in White 2005, where it's used to illustrate an explanatory virtue called *stability*. Although the point White goes on to make is different from the one I will make, and although the characterization of stability he provides is not quite the notion of counterfactual resilience that I introduce in this paper, there is much in common to the spirit of both proposals.

spend cognitive resources on details that promise little in terms of theoretical payoff. The amount of detail involved in INITIAL is unnecessary, and not because it is redundant—our preference for BIAS would remain even if we could change INITIAL so that it is silent on the bias of the coin. Rather, the explanation in terms of INITIAL is just harder to grasp than the explanation in terms of BIAS.[17]

But not all reasons for preferring BIAS over INITIAL have to do with our particular cognitive limitations. Another reason for preferring BIAS over INITIAL is that it is more general—it can be applied to many different circumstances. Generality tends to make for good explanations. This is why appealing to beliefs and desires to explain my behavior can be more satisfying than giving a full account of my brain state.[18]

Consider this example, essentially due to Alan Garfinkel.[19] Tom is running late for a meeting, because he had a leisurely breakfast. He gets in his car and drives somewhat recklessly—so much so that he loses control of the car at some point and gets into an accident. A natural explanation of this unfortunate event is that Tom was driving recklessly—that he was speeding, say. Given background assumptions, his speeding makes it very likely that he got into an accident. But this cannot be all it takes for something to be a good explanation of the accident. After all, a fuller description of Tom's morning would also make it highly likely that he got into an accident. And this, I submit, would not be a good explanation of the accident.

The reason is that, unlike the first explanation, the second one is not very portable. Had Tom not driven recklessly, we couldn't have used that as an explanation for why he got into the accident. The explanation in terms of his reckless driving, in contrast, is applicable to many other situations—there are many other ways Tom's morning could have been that, given his reckless driving, would have ended up in a car accident.

Note that both simplicity and generality have one thing in common. Having a simple, or a very general, explanation, makes the explanandum very stable.[20] The simpler the explanation, the fewer stars had to

17 Admittedly, it is tempting to think that simplicity is a virtue not just because of our cognitive limitations, but we needn't take a stance on that issue.

18 Cf. Jackson and Pettit 1988, Strevens 2004, and the discussion of causal relevance in Yablo 1992 for related discussion.

19 Garfinkel 1981, p. 30.

20 The notion of stability I am after is related to, although distinct from, the notion of

align in just the right way to make the explanandum occur. The same goes for more general explanations—the more circumstances it applies to, the easier it is that the explanandum occurs, given the explanans. This suggests a strategy for coming up with a diagnostic tool for good explanations: a way of assessing how well a putative explanation does in helping us understand the explanandum.

3.4.2 *Counterfactual resilience*

Rather than focusing on properties of the explanans, however, let us focus on what the explanans does to the explanandum, relative to a given body of beliefs. Let us first ask how well-explained a given proposition is relative to a body of beliefs. We can then use this to tell how much learning a particular proposition—a putative explanation—can contribute to having the explanandum be well-explained.

Like subjective Bayesians who think of questions of evidential support as making sense only against the backdrop of background beliefs, I think of questions of explanation as making sense only against such a background. Thus, I think it is best to ask how well explained e is relative to a given body of beliefs. Questions about how much some particular claim contributes to explaining e are, on my view, less pressing. But we can still ask whether p contributes more to an explanation of e than q does, by looking at how well-explained e would be conditional on p vs. how well-explained it would be conditional on q. And we can use this to explain why simpler, or more general, explanations are better.

My suggestion, in a nutshell, is this: the more counterfactually robust a particular claim is, the more well-explained it is (relative to a given body of beliefs). Explanations that make the explanandum less surprising, in other words, tend to be more satisfying than those that do not.[21]

One way to see that increasing the counterfactual stability of the explanandum makes for satisfying explanations is to think about laws of

---

resilience discussed in Jeffrey 1983 (when discussing the paradox of ideal evidence), or Skyrms 1977, 1980. Their focus is on stability under conditionalization—or 'indicative supposition'. Mine is on stability under counterfactual supposition.

21 Of course this will not do, as it stands, when it comes to low-probability events. But these are vexed issues far beyond the scope of this paper. See Woodward 2010 for discussion and references.

nature. Laws of nature have a high degree of counterfactual stability.[22] They are also some of the best candidates for explanatory bedrock. We all know the explanatory buck has to stop somewhere. We all agree that stopping at the laws of nature is as good a place as any. I say it is no coincidence that high counterfactual robustness goes hand in hand with not being in need of an explanation. It is because laws of nature are so counterfactually robust—because they would have obtained (almost) no matter what—that they do not cry out for explanation.

Another way of motivating the connection between counterfactual stability and explanation is to reflect on the plausibility of so-called *contrastive* accounts of explanation.[23] The idea is simple: any request for explanation takes place against the backdrop of a contrast class. What we want out of an explanation of an event *e* is a story as to why *e rather than* some other member of the contrast class occurred. Now, the harder it is to find a natural contrast class, the harder it is to reasonably expect an explanation of *e*, on this way of thinking. And if *e* has a high degree of counterfactual stability, then the harder it is to think of *e* as crying out for an explanation.

Moreover, counterfactual stability is a helpful diagnostic tool for simplicity: the fewer variables are involved in an explanation, the more robust will the explanandum be, and vice-versa. Seeking explanations that make the explanandum counterfactually robust is likely to lead to simpler explanations. And explanations that make the explanandum counterfactually robust can be applied to many different circumstances. They are 'portable', in that they can be used, *mutatis mutandis*, to explain many different phenomena.

Consider again the explanation of the sequence of nearly ten tails in a row in terms of the coin's initial conditions (together with specification of the forces involved, wind conditions, etc.)—what I called INITIAL. Slight variations in the initial conditions would have made this explanation inapplicable: there are many ways things could have been—ways similar to the way things actually are—where the explanandum might have been false.

For example, had you held the coin in a slightly different way in one

22 Indeed, some would go so far as to use counterfactual stability in order to *characterize* what laws of nature are. See, e.g. Lange 2005, 2009.

23 See Garfinkel 1981, as well as van Fraassen 1980; Lipton 1990, *inter alia*.

of the tosses, and for all the explanation tells you, the coin might have landed heads. Had someone sneezed nearby, altering the wind conditions, and the outcome might have been different. In contrast, had you held the coin slightly different, then according to BIAS the coin would have still landed tails nearly every time. BIAS is applicable to many situations—it wears its portability on its sleeve—not just involving different coins and different initial conditions, but different processes involving binary random variables.[24]

It is hard to cash the notion of counterfactual stability in a more precise way. The number of ways things might have turned out such that, for all that INITIAL says, the explanandum might have been false is infinite. But so is the number of ways things might have turned out such that, for all BIAS says, the explanandum might have been false. We cannot just *count* the relevant possibilities. And while it is in principle possible to provide a measure that would differentiate between the relevant infinite sets of possibilities, it is not obvious how to motivate one measure over another that will work for all cases.

But assume we can agree on a finite set of relevant suppositions. If the explanandum is made more robust under counterfactual suppositions in that set by one explanation than another, I submit, that would give us an *epistemic* reason (albeit a *pro tanto* one) for preferring the one over the other. This is not to say this is a reason for taking the first explanation to be more likely than the second one.[25] But you must admit that it is a reason for favoring inquiry into the truth of the one over the other.

For example, suppose you are interested in explaining why the outcome of the ten coin tosses is what it is, and in general you want to be able to explain facts about the outcome of coin tosses involving that coin. You are told your memory will be erased, but you will have some say on what credence function you will have. In particular, you are given the choice of waking up with a credence function that gets very close to the truth about the bias of the coin, and a credence function

24 There are some tricky issues I'm skating over. For example, one might think that counterfactuals of the form *If p had been false, the coin would have landed tails* cannot be true, since BIAS does not rule out entirely the possibility of the coin landing heads. For our purposes, however, these complications are best set aside.

25 Although see White 2005.

that gets very close to the truth about the initial position of each of the coin tosses.

If all else is equal, you will prefer the former over the latter. You would rather know the bias of the coin than the particular initial conditions of those ten coin tosses. After all, you can expect that having true beliefs about the bias of the coin will be more likely to explain other features of the coin. As I will put it, claims about the bias of a coin plausibly have more *explanatory potential* than claims about the particular initial conditions of some arbitrary sequence of ten coin tosses—at least relative to the sort of things we tend to want to explain. Holding fixed a class of explananda, having true beliefs that have a higher explanatory potential is, all else equal, better than not—and this, I submit, from an epistemic point of view.

Once we have an account of what the explanatory potential of a proposition is, we can make sense of the explanatory potential of a *question* as the expected explanatory potential of its correct answer. Other things being equal, the more explanatory potential a particular question has, the better it is, epistemically, to engage in finding out its correct answer. By making our epistemic utility functions sensitive to the explanatory potential of the relevant propositions, we can finally spell out a framework for evaluating different lines of inquiry from an epistemic point of view.

3.5 MEASURING EXPLANATORY POTENTIAL

Suppose we fix on a particular explanandum e. In comparing different explanations of e, I have relied on instances of following schema:

> STILL: Given the relevant explanation, if p had been false, e would have still obtained.

I want to say more about what exactly these instances are supposed to mean.

It is important to note that STILL is *not* equivalent to: had p been false, *together with the explanans* of the explanation, e would have still obtained. Given the truth of p, the closest possible world in which q is true may not be a world in which p is also true.[26]

26 I am assuming a semantics for counterfactuals broadly along the lines of Stalnaker 1968 and Lewis 1973, on which counterfactuals are sensitive to a contextually deter-

This is at it should be. For one thing that motivates the idea that BIAS is a better explanation than INITIAL is that the following is true:

> SPIN: Given BIAS, *if* the spin velocity of the coin had been different, the coin would have still landed tails nearly every time.

whereas the instance corresponding to INITIAL is not.

> SPIN-INITIAL: Given INITIAL, *if* the spin velocity of the coin had been different, the coin would have still landed tails nearly every time.

And if we understand SPIN-INITIAL as

> had the spin velocity been slightly different in any one toss and INITIAL been correct, the coin would have still landed tails nearly every time,

then we get a counterfactual with a logically impossible antecedent. After all, INITIAL specifies, inter alia, the spin velocity that each coin toss actually had. And it is hard to see why it would not be true that, had the spin velocity been both different from what it actually is and what it actually is, the coin would have landed tails nearly every time.

We should instead understand instances of STILL as follows. First, note that each explanation makes salient some relevant notion of similarity. Now take a similarity respect that is made salient by the explanation: if conditional on the explanation being true, in all the most similar worlds in which the antecedent of the counterfactual in STILL is true, e obtains, then STILL is true. The reason SPIN-INITIAL is not true is that, by slightly altering the initial conditions of the coin-tosses, we could have failed to get a sequence of nearly ten coin-tosses.

We can refine this criterion some more, by defining a degreed version of counterfactual resilience. Other things being equal, the more resilient an explanandum is according to an explanation, the better the explanation is—where resilience is understood as relative to a fixed set of candidate suppositions. Intuitively, we want to measure how stable e is under counterfactual supposition conditional on a particular explanation E. Thus, relative to a credence function C, we want to measure

---

mined similarity ordering.

the amount of variation between $C(e|E)$ and $C_\sigma(s \mathrel{\Box\!\!\rightarrow_\sigma} e|E)$, where (i) $\Box\!\!\rightarrow_\sigma$ denotes the counterfactual conditional relativized to the similarity function σ,[27] and (ii) s is an element of the fixed set of candidate suppositions.

There are different ways of measuring the relevant variation. To fix ideas, let us use a generalization of Euclidean distance. Thus, the counterfactual resilience of e, relative to a credence function C, a similarity function σ, and a finite set of suppositions S is given by the average of the square of the difference between $C(e)$ and $C(s \mathrel{\Box\!\!\rightarrow_\sigma} e)$, for $s \in S$. We can give a more explicit definition if we appeal to *imaging*[28] as a way of understanding counterfactual supposition, but those technicalities can wait for the appendix.

We can now define the counterfactual resilience of e relative to a credence function C as follows:

$$CR_{\sigma,C}(e) = 1 - 1/|S| \sum_{s \in S} (C(e) - C(s \mathrel{\Box\!\!\rightarrow_\sigma} e))^2.$$

For a given explanation E, we can now think of $CR_{\sigma,C_E}(e)$ as a measure of how robust E makes the explanandum e.

In order to build these considerations into a notion of epistemic value, we need to specify what an epistemic utility function that is sensitive to the expected degree of resilience given to an explanandum e would look like. Presumably there are different ways of bringing the expected degree of resilience to bear into a measure of epistemic utility. We have two criteria for ranking credence functions. On the one hand, we can rank them in terms of their expected accuracy. On the other, we can rank them in terms of how resilient they make a given explanandum. How exactly to weigh each of these criteria is a question for some other day.

Fortunately, we can already say something about how to compare different questions from an epistemic perspective. Recall from §3.3 that

27 One must tread carefully here. Williams forthcoming shows that, under certain assumptions, one cannot identify the credence in $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ with the credence one assigns to ψ on the counterfactual supposition that φ. For our purposes, however, we can treat $C(\varphi \mathrel{\Box\!\!\rightarrow} \psi)$ as merely short-hand for the value $C^\varphi(\psi)$—in the notation introduced below. I need not assume that there is anything as the proposition expressed by a counterfactual conditional.

28 Cf. Lewis 1976.

we had an epistemic utility function—the Brier score—on which $?p$ and $?q$ were on a par. Now, the Brier score is just a measure of expected accuracy. So it is no surprise that, from the point of view of maximizing accuracy, inquiry into p and inquiry into q are taken to be on a par.[29] But we can use counterfactual resilience as a tie breaker. In other words, we can add one more constraint on epistemic utility functions to our list:

> RESILIENCE: If all else is equal, and the expected resilience of e relative to C and S is higher than the expected resilience of e relative to C' and S', then the expected epistemic utility of C should be higher than the expected epistemic utility of C'.

Go back to the example of the explanations of the sequence of ten tails. From the point of view of maximizing accuracy, inquiry into BIAS and inquiry into INITIAL may well be on a par. It will depend on the prior degrees of belief you assign to the relevant propositions. But from the point of view of explanatory potential, I have argued, inquiry into BIAS is to be preferred.

3.6 IMMODESTY AND EPISTEMIC IMAGINATION

Note that what we have so far is a way of comparing different lines of inquiry from an epistemic point of view. We have, in other words, an account that would allow us to tell which questions we should try to answer from an epistemic perspective. If you expect that answering Q is more likely to give you an explanation of what you want to explain than answering Q', then you should prefer to gather evidence that bears on Q rather than Q'.

Thus, nothing in what I've said so far suggests that Q having a higher expected epistemic value than Q' should in any way affect your epistemic state.[30] Yet if assessments of the epistemic value of different lines of inquiry are to be incorporated into an account of rational inquiry, it seems like such assessments *should* be able to change, in some cases, your epistemic state.

---

29 This brings out the fact that we have not taken into account how likely we take finding out the answer to question to be.

30 At least when it comes to propositions that are not about the difference in expected epistemic value of the relevant questions.

One reason to think this cannot be is that realizing that Q has higher expected epistemic value than Q' does not involve acquiring new evidence. And in order for a change in our epistemic state to be *rational*, it would seem, it must be triggered by the acquisition of new evidence. After all, epistemic rationality requires a certain amount of immodesty. If you are rational, then you should take yourself to be doing as well as you can, *epistemically*, given your available evidence. To change your epistemic state without new evidence would involve moving to an epistemic state that is *worse* than the one you currently are.

Appearances are misleading, however. One can consistently maintain that epistemic rationality requires that we take ourselves to be doing as well as we can, epistemically, and that some rational epistemic changes need not be triggered by new evidence. To see what I have in mind, consider the following analogy.

Lars is a *fun-maximizer*. At any point in time, he always takes himself to be doing the most fun thing he can do. While having lunch, Lars' perspective on the world is quite striking: I could be doing a number of things right now, but nothing would be as fun as having lunch. Of course, Lars is not stuck having lunch all day long. As he eats, his evidence changes: he acquires evidence that he is no longer hungry, and proceeds to do what he takes to be the most fun activity he could do, in light of his evidence.

If you looked at the life Lars lives, you would find it incredibly boring. Not because his preferences are much different from yours. Rather, the problem is that Lars lacks imagination. If you could only get him to see that there are many things he could do with his day other than spend it quietly under an oak tree, you know he would be thankful. And this can be so even thought Lars is doing as well as he can in order to have fun. Among all the options that ever occur to him as things he could be doing at any particular time, he always takes himself to be doing the one he finds most fun. But this is not because he is always having that much fun. Rather, it's because his lack of imagination precludes him from seeing all the fun things he could do instead.

Now consider an agent, Tom, who always takes himself to be doing as well as he can, epistemically. He only has beliefs about a particular issue: whether he is tired. He is very good at responding to the evidence he receives, and at any time, he takes himself to be doing as well as he can with regards to the issue of whether he is tired. To some extent, Tom

is doing quite well, epistemically. But he could be doing much better: he could be asking questions about many things, including issues that have little to do with how tired he is.

The point is a rather simple one. If Tom, like Lars, lacks imagination, he could take himself to be doing as well as he can, epistemically, because he only considers a limited range of options. If he came to see that he could be in an epistemic state he had not considered, he would pick it in a heartbeat. So if a new issue occurred to him, he could in principle come to have a view on the matter without having acquired any new evidence. And crucially, without taking his earlier self to have been at fault.

This a type of epistemic change that orthodox Bayesian theories of epistemic rationality have little to say about. I want to sketch a way of thinking about epistemic rationality that is conservative, in that it allows us to capture some basic principles behind traditional Bayesian accounts of epistemic rationality. But it gives us the flexibility to ask questions about the rationality of conceptual change: about how epistemic rationality could constrain expansions of the range of hypothesis we rely on in inquiry.

3.7 RATIONAL DYNAMICS AND EPISTEMIC VALUE

If we make some plausible assumptions, claims about epistemic value can be used to motivate the standard norms of Bayesian rationality. For instance, Joyce (1998, 2009) shows that on certain plausible assumptions about epistemic utility functions, probabilistically incoherent credences are dominated by probabilistically coherent ones: if your credences are not probabilistically coherent, there will be some probability functions that you take to have higher expected epistemic value. And Greaves and Wallace (2006) show that, once an agent receives evidence E, updating by conditionalization on E is what maximizes expected epistemic value—again, assuming that epistemic utility functions are proper, and that your prior degrees of belief are probabilistically coherent.

But an account of epistemic rationality in terms of epistemic value gives us more flexibility. Suppose our agent, Tom, has never considered a particular question Q. Further suppose that, once the question occurs to him, he can estimate a particular expected epistemic value to learning

the answer to Q. Then it might be rational for him to gather evidence that bears on Q: and this, I submit, will require a certain change in Tom's epistemic state, one that is not recognized by any broadly Bayesian theory of epistemic rationality. Let me explain.

I have been thinking of an agent's epistemic state as a credence function: an assignment of numerical values to a set of propositions that satisfies the standard Kolmogorov axioms of the probability calculus.[31] If a proposition p is in the domain of an agent's credence function C, then $C(p)$ represents the degree to which the agent believes p. What exactly this amounts to needs not concern us for now. The basic thought is that we can recover comparative judgments of likelihood from a credence function. In short, what matters is that, e.g. if an agent takes p to be more likely (in an epistemic sense) than q, then $C(p) > C(q)$.[32]

Crucially, a credence function assigns value to some set of propositions: it needn't assign numerical values to *all* propositions. Why not take an agent's epistemic state to assign a numerical value to all propositions? Of course, if we are not interested in what value an agent assigns to a particular collection of propositions, we could simply have them drop out of the picture. But in that case, a proposition not being in the domain of an agent's credence function would not tell us anything of substance about the agent's epistemic state. It would simply tell us that we are not interested in what value the agent assigns to that proposition. Is there anything we could model about the agent by leaving a proposition out of the domain of her credence function?

Start out with an easy case. Tom is a much more primitive version of Jackson's Mary.[33] He lives in a black and white world but he knows very little physics. Further, he has not even heard of color words. He is unable to distinguish red things from the rest not just visually, but in any other way.

It is quite tempting to say that Tom has no doxastic attitude towards propositions about the colors of things. In particular, the proposition

31 For the sake of mathematical expediency, I am assuming that the domain of a credence function is a field of sets.

32 One thing to note is that, by requiring that Cr be a probability function, we are unable to model cases in which an agent takes neither p nor q to be more likely than the other, nor does she take them to be equally likely. This is simply because the natural order of real numbers is a total order.

33 Jackson 1986.

that he is wearing all blue is one he has no view on. But it is also tempting to say something stronger, viz. that Tom is not even aware of that proposition: given his current epistemic state, there is little sense to be made of him suspending judgment on that proposition. His epistemic state is blind to that proposition, as it were. In general, when an agent is unable to entertain a given proposition, there is a principled reason for leaving it out of the domain of her credence function.[34]

I suspect you are still on board. Nothing I've said about what partial credences can be used to model is terribly controversial. But neither is it terribly interesting. After all, it is roughly a psychological question what sort of distinctions an agent is able to make. And whereas there surely are some interesting questions about the way in which Tom could come to acquire the relevant distinctions, they are better left for those without enough budget to run a lab.

Yet there is a further distinction to be made here. Tom, we suppose, has no doxastic attitudes towards propositions about the color of things. The reason is that, as described, Tom lacks the conceptual resources to make the relevant distinctions. But suppose Tom acquires the ability to make these distinctions. Should we insist that Tom's epistemic state be modeled by a credence function that assigns value to all propositions about the color of things?

From our point of view, as modelers, there is little to recommend this. It is hard to see what reasons we would have for using some real number or other to represent Tom's attitude to the proposition that some potatoes are blue. But even from *Tom's* point of view, the proposition that some potatoes are blue is not even part of his epistemic landscape. There is a sense in which Tom will, and in my view *should*, ignore this proposition. Having his credence function be undefined on that proposition is a way of modeling just that.[35]

Of course, things can change. It can be that it will become advantageous for Tom to gather evidence bearing on the question of whether some potatoes are blue. And in order to make room for this in our

34 Note that the suggestion so far is not to leave out a proposition p whenever the agent is not attending to the proposition. We may good grounds for thinking that the agent assigns a particular credence to p even if the agent is not currently entertaining the issue of whether p.

35 Cf. Rayo 2011.

model, we need to allow for changes in his epistemic state that involve the *expansion* of the domain of his credence function.

Bayesian epistemologists tend to focus on a distinctive type of epistemic change: when an agent's credence function comes to assign different values to a given body of propositions. Some of these changes are said to be rational—those resulting from conditionalization on the evidence available to the agent, say. Some are not: my moving from uncertainty as to whether p to full certainty as to whether p without having acquired any new evidence. But Bayesians have nothing to say about expansions: the orthodox Bayesian machinery lacks the resources to even ask whether some expansions could be epistemically rational.[36]

In contrast, by moving to a framework in which rational dynamics involves maximizing expected epistemic value, we can ask the question of whether a particular expansion is epistemically rational. In particular, we can ask the question of whether introducing new conceptual machinery, or postulating new hypotheses, are likely to be beneficial, epistemically.

3.8 THE RATIONALITY OF CONCEPTUAL CHANGE

Start with a simple toy example.[37] You are studying an unfamiliar type of organism, call them 'Reds', and their reactions to certain stimuli. You keep your Reds inside dark boxes for a little while and then proceed to flash different colored lights on them to see how they react.

You notice that Reds that were exposed to red light tend to stop moving altogether until the lights are switched off. In contrast, exposing Reds to blue or green light seems to have little or no effect on their behavior.

However, if you take a Red that was previously exposed to red light, you observe that exposing it to blue light tends to make it move significantly faster than normal. It occurs to you that there could be in an

36 You might think that PROPRIETY should be enough to rule out this possibility, once we move to a framework in which rational dynamics is a matter of maximizing expected epistemic value. But as I show in the Appendix, PROPRIETY is only a reasonable constraint when we are focusing on credence functions that assign values to the same collection of propositions.

37 This example is based on a series of cases discussed in great detail in Sober 1998. See also Forster 1999, for related discussion of how conceptual innovation can be motivated by epistemic considerations.

internal state *R* such that a Red could get in state *R* as a result of being exposed to red light, and once in state *R* it would respond to blue light differently than it would had it not been in state *R*.

Once you bring *R* into the picture, you can formulate a hypothesis about Reds that could be used to explain why Reds would respond to blue light by moving faster after being exposed to red light. For example, that Reds in state *R* tend to get excited by blue lights, and that exposure to red light tends to cause Reds to be in state *R*. Under this hypothesis, the claim that a particular Red will move faster when exposed to blue light is made more counterfactually robust than it would be otherwise. For given that they are in state *R*, had they not been exposed to Red light, they would have still responded to blue light the way they did (given the new hypothesis). In other words, the introduction of state *R* allows for the formulation of a hypothesis that would, *if true*, increase the counterfactual resilience of the claim that a given Red would respond the way it did to blue light.

Now, nothing here suggests that you should conclude that Reds do *in fact* get to be in state *R* when exposed to red light. But the expansion of your hypothesis space to include that particular hypothesis—which will have to be assessed in light of the data at hand[38]—can be motivated by the considerations of epistemic utility above. After all, some claims involving state *R* have a high explanatory potential, so the expected epistemic value of expanding your credence function to allow for such propositions is relatively high.[39]

Here is a slightly more complex example, due in its essentials to Frank Arntzenius.[40] You arrive in a strange land, where you find a collection of round critters, each of about 1 inch in diameter. Each critter is either red or white. You pick some up and discover that, if you press two critters against each other, they combine to form a larger critter, of about 2-inches in diameter, uniform in color—red or white. You try combining two 2-inch critters and discover that they too combine to form a larger critter, of about 3-inches in diameter, uniform in color.

38 On which, again, see Sober 1998. Specifically, Sober discusses ways in which the introduction of intervening variables can sometimes be motivated by frequency data.

39 See the appendix for an illustration of how one can compute the expected epistemic value of a given expansion even though one hasn't assigned credence to the 'new' propositions.

40 Arntzenius 1995.

You set out to understand how the colors of smaller critters relate to the color of the larger critters they turn into when combined. After gathering data for a while, you have the following observations. First, if you combine two 1-inch critters, they will turn into a 2-inch *red* critter unless both the 1-inch critters were white. If you combine two 2-inch critters, however, things get slightly trickier.

If you combine two 2-inch white critters, they combine to form a white 3-inch critter. But if at least one of the two 2-inch critters is red, the color of the resulting 3-inch critter is sometimes white and sometimes red. After trying this out with a large number of 2-inch critters, you get the following frequencies. First, if one of the 2-inch critters is red, and it came from two 1-inch red critters, then no-matter what other 2-inch critter it's combined with, the result will be a 3-inch critter. But if you only look at combinations of 'mixed' 2-inch red critter with other 2-inch critters—either mixed red or white—the color of the resulting 3-inch critter will have the following distribution:

	3-inch red	3-inch white
mixed 2-inch red	75%	25 %
2-inch white	50%	50%

Now, imagine you want to explain why a particular 3-inch critter is red. You look at your records and notice that it came from two 2-inch red critters. How good of an explanation is this? Granted, on the basis of your observation, your credence that a 3-inch critter will be red given that it came from two 2-inch red critters is relatively high. But this explanation does not make your explanans particularly robust. For example had one of the red 2-inch critters come from different colored parents, they might have combined to form a white critter instead. So the color of the 2-inch critters does not, by itself, suffice to make the explanandum counterfactually robust.

Suppose now it occurs to you that the red critters could come in two varieties—strong-red and weak-red.[41] If a 2-inch red critter comes from two 1-inch red critters, it is strong-red. If it comes from a 'mixed'

41 Of course, this is just a simplified version of the conceptual innovation behind Mendelian genetics.

pair of critters, it is weak-red.[42] You now note that a 2-inch strong-red critter combined with a 2-inch white critter yields a 3-inch red critter, and that a 2-inch weak-red critter combined with a 3-inch white critter yields a 3-inch red critter 50% of the time, and a 3-inch white critter 50% of the time. You can now formulate the following hypothesis—a hypothesis that was not part of your hypothesis state before you considered that red critters could come in two types:

	s-red	w-red	white
s-red	s-red	50% s-red 50% w-red	w-red
w-red	50% s-red 50% w-red	25% s-red 50% w-red 25% white	50% w-red 50% white
white	w-red	50% w-red 50% white	white

If you were to add this hypothesis to your body of beliefs, you could make claims about the heredity of color features among the critters more robust: assuming that a red critter is strong-red you could now explain why it would, when combined with a blue critter, yield a red critter most of the time. Knowing what type of red a critter is would allow you to explain things about the distribution of color among its offspring independently of what generation the critter happens to be—you've thereby made your explanandum resilient under the counterfactual supposition that the given red critter is a third-generation, say, rather than a second-generation one—and also independently of what its 'parents' were—you've made your explanandum resilient under the assumption that your critter had different colored ancestors, say.

Now, in both these cases, conceptual innovation allowed for the formulation of hypotheses that were not part of the starting hypothesis space. There is still a question to be asked, viz. how is it that a given hypothesis gets entertained for the first time? But perhaps there isn't much to be explained here.

42 The sense in which these are 'new' properties is this: these properties are causally dependent on, but they are not reducible to, facts about the critters' lineage.

Imagine a machine that is generating possible new hypothesis at random—new ways of partitioning the state space it is working on. Allow for the machine to evaluate each possible hypothesis in terms of its expected epistemic value. By constraining the process of crafting its hypothesis space by considerations of epistemic value the machine is more likely to yield better theories. Rather than undertaking avenues of inquiry that would lead to nowhere. Not because theories with high expected epistemic value are more likely to be true—but because *if true*, they are more likely to be more satisfying, from an epistemic point of view.[43]

3.9 CONCLUSION

Let me take stock. I started out with an intuition: that some lines of inquiry are better, *epistemically*, than others. I proposed a way of cashing out this intuition: by assigning epistemic value to different bodies of belief, we can evaluate a given line of inquiry on the basis of the value of the body of belief that this line of inquiry is expected to yield. I outlined a framework for understanding this notion of epistemic value, by looking at the extent to which a given body of beliefs was explanatorily closed. I then conjectured that counterfactual resilience could be a tractable guide to explanatory closure.

This strategy had two additional benefits. First, it allows us to assess *expansions* of our hypothesis space before setting off to gather evidence for the new hypothesis. We should only spend cognitive resources on new lines of inquiry that promise to be epistemically valuable. Of course, having a high expected epistemic value is no guarantee that the given line of inquiry will prove to be helpful. But it gives us an epistemic reason to look into it—to take the relevant hypotheses seriously in inquiry. This strategy thus provides us with a model of how conceptual change and theoretical innovation could fall under the scope of a theory of epistemic rationally. How much this is so remains to be seen. But at the very least, it gives us a framework for asking questions about the rationality of a type of epistemic change that was ruled out by default from an orthodox Bayesian framework.

43 Cf. Bromberger 1992 for more on the role of questions in, and the important of formulating new questions for, inquiry.

3.A APPENDIX

I will outline a formal framework for investigating how epistemic utility functions can be used to assess different expansions of a credence function. But first, some definitions are in order.

3.A.1 *Basic definitions*

Given any probability function P, let $\mathcal{A}_P$ denote the domain of P. To simplify our discussion, we will restrict our attention to atomic algebras. Thus, for each P, we can define π_P as the smallest subset of $\mathcal{A}_P$ whose Boolean closure is $\mathcal{A}_P$. Note that π_P is a partition of $\mathcal{W}$.

For our purposes *utility function* is a function u that associates, to each probability function P and world $w \in \mathcal{W}$ a real number $u(P, w)$, which is the *score* of P in w.

Intuitively, any such function must satisfy the following desideratum: if P does not distinguish between w and w', then $u(P, w) = u(P, w')$. Thus, if w is not in the domain of P, then $u(P, \cdot)$ will be constant throughout the π_P cell of w. The following definition captures this intuition:

Definition 3.A.1: A utility function u is *nice* iff for each P, $u(P, \cdot) : \mathcal{W} \to \mathbb{R}$ is P-measurable.

From now on, I will assume that all utility functions are nice.

If a probability function P is defined over the entire power set of $\mathcal{W}$, then we can define the *expected score* of any probability Q relative to u in the usual way, viz.

$$EU_{u,P}(Q) = \sum_{w \in \mathcal{W}} P(w)u(Q, w).$$

But we need a different definition in order to allow for π_P to be coarser than the set of singletons of $\mathcal{W}$. The above definition is of no help, since there will be some $w \in \mathcal{W}$ such that $\{w\} \notin \pi_P$, and thus $P(w)$ will be undefined.

The best we can do is to approximate the expected value of Q relative to u and any *extension* of P to the entire power set (at least to π_Q—as we will see, this makes no difference under the assumption that u is nice).

For any algebra $\mathcal{A}$ and any probability function P such that $\pi_P \subset \mathcal{A}$, let $\mathbb{P}_P(\mathcal{A})$ denote the set of all extensions of P to $\mathcal{A}$. If Q is a probability

function, I will write $\mathbb{P}_P(Q)$ to denote $\mathbb{P}_P(\pi_Q)$. I will use $\mathbb{P}_P$ to denote the set of all extensions of P to the entire power set.

Fix a probability function P. We can now define, for each probability function Q and each $s \subset \mathcal{W}$

Definition 3.A.2:

$$\overline{EU}(Q,s) = _{P' \in \mathbb{P}_P} \sum P'(w|s)u(Q,w)$$
$$\underline{EU}(Q,s) = _{P' \in \mathbb{P}_P} \sum P'(w|s)u(Q,w)$$

Since u is nice, whenever $w \in q \in \pi_Q$ we have $\overline{EU}(Q,q) = \underline{EU}(Q,q) = u(Q,w)$, so we can extend our function so that $u(Q,q)$ is well-defined. This has as an immediate consequence the following easy fact:

Fact 3.A.3: For any $P' \in \mathbb{P}_P$, and any Q,

$$\sum P'(w|s)u(Q,w) = \sum_{q \in \pi_Q} P'(q|s)u(Q,q).$$

We can now define the upper and lower expected values of an extension Q of P as follows:

Definition 3.A.4:

$$\overline{EU}(Q) = \overline{EU}(Q,\mathcal{W}) = _{P' \in \mathbb{P}_P(Q)} \sum_q P'(q)u(Q,q).$$
$$\underline{EU}(Q) = \underline{EU}(Q,\mathcal{W}) = _{P' \in \mathbb{P}_P(Q)} \sum_q P'(q)u(Q,q).$$

Clearly, $\overline{EU}(Q) \geq \underline{EU}(Q)$, with equality iff $\pi_P = \pi_Q$.[44]

Now, given a utility function u, we can compare two extensions of Q, Q' of P in many ways. For example, we can ask which one maximizes $\overline{EU}$, which one minimizes $\underline{EU}$, etc. Presumably there will be things to be said in favor of each of these decision rules. But we need a better understanding of what utility functions are like if these decision rules are not to degenerate into triviality. Further, we need to see whether there are any useful generalizations to be made given a set of constrains on utility functions.

44 Note that if Q is a coarsening of P, then $EU_{P,u}(Q) = \sum_p P(p)u(Q,p)$ is well-defined.

3.A.2 *Epistemic utility functions*

There is a substantial body of literature on so-called *scoring rules* or *epistemic utility functions.*[45] All extant discussions, however, restrict their attention to functions of the form

$$u : \mathbb{P}(\mathcal{A}) \times \mathcal{W} \to \mathbb{R},$$

where $\mathbb{P}(\mathcal{A})$ is the set of all probability distributions over a fixed algebra $\mathcal{A}$. In this context, epistemic utility functions are utility functions that satisfy a number of constraints, like TRUTH-DIRECTEDNESS, or PROPRIETY. How far can we generalize these constraints to the case at hand? In other words, how should we state versions of these constraints for epistemic utility functions whose domain includes pairs of the form (P, w) and (Q, w) with $\pi_P \neq \pi_Q$?

The weakest extension of these principles would just require that an epistemic utility function satisfies TRUTH-DIRECTEDNESS and PROPRIETY when restricted to a given algebra. On my view, this is the only plausible generalization to epistemic utility functions that can evaluate probability distributions over different domains. Let me explain

Let us first consider TRUTH-DIRECTEDNESS. The only candidate extension that seems to make sense would be this. First, assume that π_P and π_Q have the same cardinality. Fix a bijection $f : \pi_P \to \pi_Q$. Then the generalized version of TRUTH-DIRECTEDNESS would require that if for all $s \in \pi_P$, $|P(s) - s(w)| \leq |Q(f(s)) - f(s)(w)|$, then $u(P, w) \geq u(Q, w)$. The problem is to find a principled way of fixing the bijection. For a given $w \in \mathcal{W}$, we can require that $s(w) = f(s)(w)$, but this does not give us that much traction. If we had some version of EXTENSIONALITY (Joyce, 2009) then perhaps we could motivate this way of generalizing the principle.

Now, the case against generalizing PROPRIETY is more straightforward. Again, the minimal change we need to make to PROPRIETY is what I'll call PARTITION-WISE PROPRIETY, which essentially amounts to the claim that $u \restriction \mathbb{P}(\mathcal{A}) \times \mathcal{W}$ must be proper. But beyond this, the only generalization that can be motivated is this:

UNIVERSAL PROPRIETY: For any $P \neq Q$, $EU_{P,u}(P) > \overline{EU}_{P,u}(Q)$.

45 E.g. Greaves and Wallace 2006; Joyce 1998, 2009.

Unfortunately, UNIVERSAL PROPRIETY cannot be satisfied by *any* epistemic utility function:

Fact 3.A.5: No nice epistemic utility function can be universally proper.

Proof. Assume otherwise and let u be universally proper, and let Q be a non-trivial extension of P. Since $Q \neq P$, we have

$$\sum_q Q(q)u(Q,q) > \sum_q Q(q)u(P,q) = \sum_p Q(p)u(P,p).$$

where the last equality follows from the probability calculus.[46]

Now, since Q is an extension of Q, we have $P(p) = Q(p)$ for $p \in \pi_P$. Thus, since u is universally proper, we have:

$$\sum_q Q(q)u(Q,q) > \sum_p P(p)u(P,p) >$$
$$P' \in \mathbb{P}_P(Q) \sum_q P'(q)u(Q,q) \geq \sum_q Q(q)u(Q,q).$$

a contradiction. □

The notion of counterfactual resilience promises to give us a way of extending our framework in order to compare probability functions with different domains.

3.A.3 *Counterfactual resilience*

Let us fix a class S of *potential suppositions*. The degree of counterfactual resilience of e, relative to P and S, is given by:

$$CR_{S,P}(e) = 1 - \sum_{s \in S}(P(e) - P(s \boxdot\!\!\rightarrow e))^2.$$

In other words the more counterfactually resilient e is, relative to P and S, the more robust the value of $P(e)$ is under counterfactual suppositions with elements of S.

Now, suppose we have a class E of *explananda*. Other things being equal, a probability function that assigns to each $e \in E$ a high degree of

46 Since u is nice, $u(P,q) = u(P,\pi_P(q))$, where π_P is the projection onto π_P of q, so $u(P,q)$ is well-defined.

counterfactual resilience relative to S is better, epistemically, than one that does not. What we want is for our epistemic utility functions to track these differences.

To get there, however, we need to say something about how $P(s \boxdot\!\!\rightarrow e)$ is related to $P(s)$ and $P(e)$. (This is because we have no guarantee that if s and e have a well-defined value under P, so does $s \boxdot\!\!\rightarrow e$.) In other words, we need to say something about the probabilities on counterfactuals: better yet, about credences in counterfactuals.[47]

3.A.4 *Generalized imaging*

Fix $P : \mathcal{A} \to [0, 1]$ a probability function, and let $\mu_c : \mathcal{A} \times \pi_P \to [0, 1]$ be a probability functions for each $c \in \mathcal{A}$.

Definition 3.A.6: The *image* of P on c relative to μ is defined as:

$$P_\mu(x\backslash c) = \sum_{s \in \pi_P} P(s)\mu_c(x, s)$$

Intuitively, μ_c is supposed to correspond to a measure of similarity among worlds: $\mu_c(x, s)$ tells you how s fares in terms of closeness to x among worlds in which c holds.

Following Joyce and Lewis,[48] I will take as primitive a similarity function such that $s[a]$ consists of those a-worlds that are most similar to s. Given this similarity function, we can define our similarity measure as follows

$$\mu(x, s, c) = P(x|s[c]),$$

so that

$$P_\mu(x\backslash c) = \sum_{s \in \pi_P} P(s)P(x|s[c])$$

Now, presumably, we want $P(x\backslash c)$ to be defined even when the ratio $P(x \wedge c)/P(c)$ is not (either because $P(c)$ is undefined, or because $P(c)$

47 There is a long tradition in philosophy linking beliefs in conditionals with conditional probabilities. The idea goes back to Ramsey, and the pipe dream is to find a semantics for the indicative conditional '>' such that $P(a > b) = P(b|a)$. It is well-known that this is just that: a pipe dream. In light of a number of triviality results, we now know that there is no interesting way of designing such a semantics. Despite all this, it is widely acknowledged that there is a deep connection between our beliefs in indicative conditionals and the corresponding conditional beliefs.

48 Cf. Lewis 1976. I will be using the particular formulation due to Joyce 1999.

is zero), so we will need to stipulate that conditional probabilities are primitive.

Remark 3.A.7: *Assume the $s[c]$ form a partition of c; further assume that $P(x|s[c])$ is defined for each $s \in \pi_P$, $c \in \mathcal{A}$. Then it follows from the probability calculus that*

$$P(x|c) = \sum_s P(s[c]|c)P(x|s[c]).$$

Thus, both imaging and conditionalization are weighted averages of $P(x|s[c])$, but the weights will differ in general.

3.A.5 *Imaging and expansions*

I have left implicit the dependence of $P(x\backslash c)$ on the similarity function $(s, c) \mapsto s[c]$, but it is time to make it explicit. From now on, I will write $P_\sigma(\cdot\backslash c)$ to denote the image of P on c relative to σ, where $\sigma : \pi_P \times \mathcal{A} \to \mathcal{A}$. I will also denote by $s_\sigma[c]$ the set of c-worlds that are most similar to s relative to σ.

Note that in order the image of P on a condition c to be well-defined, relative to a similarity function σ, on a point x, all the relevant conditional probabilities of the form $P(x|s_\sigma[c])$ for each $s \in \pi_P$ must be well-defined. This gives us a notion of *accessibility* for probability functions:

Definition 3.A.8: A probability function P has *access* to a similarity function σ (relative to a set of suppositions S, and and explanandum e) just in case, for all $s \in \pi_P$, and all $c \in S$, $P(e|s_\sigma[c])$ is well-defined.

The reason this is relevant to our purposes is that if P is an expansion of Q (that is, $\pi_Q \subsetneq \pi_P$, and for each $q \in \pi_Q$, $P(q) = Q(q)$), P may have access to more similarity functions than Q. This because even when $Q(s)$ is well-defined, $Q(x|s_\sigma[c])$ may not be.

3.A.6 *Toy models*

I now want to revisit the first toy example of §3.8. Why does the postulation of a new variable—reflecting whether a Red was in state R—increase the expected counterfactual resilience of our explanandum?

Let C be your credence function before the postulation of the new variable. Let red-n stand for the proposition that a given Red was exposed to red light at time t_n (similarly for blue, and green). Let faster

stand for the proposition that a given Red moves faster than normal. The following credence assignments could represent your degrees of belief: $C(\text{red-1}) = 1$, $C(\text{faster} \mid \text{red-1}) = 0$, $C(\text{faster} \mid \text{blue-1}) = 0.1$, $C(\text{faster} \mid \text{blue-2}) = .3$, $C(\text{faster} \mid \text{blue-2}, \text{red-1}) = .9$.

Let e be the explanandum: that the Red is moving faster than normal at time t_2. Presumably, after your observations, $C(e)$ is quite high. Nevertheless, if we let S contain all the descriptions of possible light colors the given Red could have been exposed to at time t_1, we have that $CR_{S,C}(e)$ is not too high. (For $C(\text{blue-1} \mathrel{\square\!\!\rightarrow} e)$ and $C(\text{blue-1} \mathrel{\square\!\!\rightarrow} e)$ are much lower than $C(e)$.)

Now, consider the question whether Red responds to blue light the way it does by virtue of being in state R. Call this proposition H. You have no well-defined credence over H. Nevertheless, the counterfactual resilience of e relative to C_H is high, using as a similarity function the partition generated by R—two worlds are equivalent just in case they agree on whether Red is in state R. This is because for all $s \in S$:

$$C_H(e\backslash s) - C(e) = C_H(R)C(e|R) - C(e) \approx 0.$$

To assess the value of $?H$, of course, we also need to estimate the resilience of e relative to $C_{\neg H}$. But this will presumably be equal to the prior resilience of e. Thus, learning the answer to $?H$ can be expected to increase the resilience of e.

BIBLIOGRAPHY

Arntzenius, F., 1995: A Heuristic for Conceptual Change. *Philosophy of Science*, 62(3): pp. 357–369.

———, 2008: Rationality and Self-Confidence. In *Oxford Studies in Epistemology*, T.S. Gendler and J. Hawthorne, eds., vol. 2, pp. 165–178. Oxford: Oxford University Press.

Benacerraf, P., 1973: Mathematical Truth. *The Journal of Philosophy*, 70(19): pp. 661–679.

Berkeley, G., 1732: *Alciphron: Or the Minute Philosopher*. In *The Works of George Berkeley, Bishop of Cloyne*, A. Luce and T. Jessop, eds., vol. III. Edinburgh: Thomas Nelson. 9 volumes. 1948–1957.

Blackburn, S., 1990: Hume and Thick Connexions. *Philosophy and Phenomenological Research*, 50: pp. 237–250.

———, 1998: *Ruling Passions*. Oxford University Press.

Brier, G., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): pp. 1–3.

Bromberger, S., 1966: Why-Questions. In *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, R.G. Colodny, ed., vol. 3 of *University of Pittsburgh Series in the Philosophy of Science*, pp. 86–111. Pittsburgh, PA: University of Pittsburgh Press.

———, 1988: Rational ignorance. *Synthese*, 74(1): pp. 47–64.

———, 1992: *On What We Know We Don't Know*. Chicago & Stanford: The University of Chicago Press & CSLI.

Broome, J., 2002: Practical reasoning. In *Reason and nature: Essays in the theory of rationality*, J. Bermúdez and A. Millar, eds., pp. 85–111. Oxford: Oxford University Press.

Burgess, J.P. and Rosen, G., 1997: *A Subject with No Object: strategies for nominalistic interpretation of mathematics*. New York: Oxford University Press.

Chalmers, D.J., 2003: Consciousness and its place in nature. In *The Blackwell guide to the philosophy of mind*, S.P. Stich and T.A. Warfield, eds., pp. 102–142. Oxford: Blackwell.

Detlefsen, M., 2005: Formalism. In *The Oxford Handbook of Philosophy of Mathematics and Logic*, S. Shapiro, ed., pp. 236–317. Oxford: Oxford University Press.

Dreier, J., 2006a: Negation for Expressivists: A Collection of Problems with a Suggestion for their Solution. *Oxford studies in metaethics*, 1: pp. 217–233.

———, 2006b: Disagreeing (about) What to Do: Negation and Completeness in Gibbard's Norm-Expressivism. *Philosophy and Phenomenological Research*, 72(3): pp. 714–721.

Dretske, F., 1981: *Knowledge and the Flow of Information*. Cambridge, Mass.: MIT Press.

Evans, G., 1982: *The Varieties of Reference*. Oxford: Clarendon Press.

Field, H., 1980: *Science Without Numbers: A Defence of Nominalism*. Princeton: Princeton University Press.

———, 1982: Realism and anti-realism about mathematics. *Philosophical Topics*, 13: pp. 45–69. Reprinted in (Field, 1989).

———, 1989: *Realism, Mathematics, and Modality*. Oxford: Basil Blackwell.

———, 1990: 'Narrow' Aspects of Intentionality and the Information-Theoretic approach to Content. In *Information, Semantics and Epistemology*, E. Villanueva, ed., pp. 102–116. Oxford: Basil Blackwell.

———, 2001: *Truth and the Absence of Fact*. New York: Oxford University Press.

———, 2009: Epistemology without metaphysics. *Philosophical Studies*, 143(2): pp. 249–290.

———, 1978: Mental representation. *Erkenntnis*, 13(1): pp. 9–61. Reprinted in (Field, 2001).

Fodor, J., 1975: *The language of thought*. Cambridge, Mass.: Harvard Univ Press.

Forster, M.R., 1999: How Do Simple Rules 'Fit to Reality' in a Complex

World? *Minds and Machines*, 9(4): pp. 543–564.

van Fraassen, B.C., 1980: *The Scientific Image*. Oxford University Press.

Frege, G., 1884: *Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl*. Breslau: Willhelm Koebner. Trans. by J. L. Austin as *The Foundations of Arithmetic: A logico-mathematical enquiry into the concept of number*. Oxford: Basil Blackwell, 1980.

Garfinkel, A., 1981: *Forms of Explanation*. New Haven: Yale University Press.

Gibbard, A., 2008: Rational credence and the value of truth. In *Oxford Studies in Epistemology*, T.S. Gendler and J. Hawthorne, eds., vol. 2, pp. 143–164. Oxford: Oxford University Press.

———, 1990: *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, Mass.: Harvard University Press.

———, 2003: *Thinking How to Live*. Cambridge, Mass.: Harvard University Press.

Gödel, K., 1947: What is Cantor's Continuum Problem? *The American Mathematical Monthly*, 54(9): pp. 515–525.

Good, I.J., 1967: On the Principle of Total Evidence. *The British Journal for the Philosophy of Science*, 17(4): pp. 319–321.

Greaves, H. and Wallace, D., 2006: Justifying conditionalization: Conditionalization maximizes expected epistemic utility. *Mind*, 115(459): pp. 607–632.

Hare, R., 1952: *The language of morals*. Oxford: Clarendon Press.

Harman, G., 1977: *The Nature of Morality: An Introduction to Ethics*. Oxford University Press.

———, 1986: *Change in view*. Cambridge, Mass.: The MIT Press.

Jackson, F., 1986: What Mary Didn't Know. *Journal of Philosophy*, 83(May): pp. 291–295.

Jackson, F. and Pettit, P., 1988: Functionalism and Broad Content. *Mind*, 97(387): pp. 381–400.

Jeffrey, R.C., 1983: *The Logic of Decision*. University Of Chicago Press.

Joyce, J.M., 1998: A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65(4): pp. 575–603.

———, 1999: *The Foundations of Causal Decision Theory*. New York: Cambridge Univ Press.

———, 2009: Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In *Degrees of Belief*, F. Huber and C. Schmidt-Petri, eds., vol. 342, chap. 11, pp. 263–297. Dordrecht: Springer Netherlands.

Lange, M., 2005: Laws and their stability. *Synthese*, 144(3): pp. 415–432.

———, 2009: *Laws and Lawmakers*. New York: Oxford University Press.

Leuenberger, S., 2004: Humean Humility? Unpublished manuscript, Australian National University.

Lewis, D.K., 1973: *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

———, 1976: Probabilities of conditionals and conditional probabilities. *The Philosophical Review*, 85(3): pp. 297–315.

———, 1979a: Attitudes *De Dicto* and *De Se*. *The Philosophical Review*, 88(4): pp. 513–543.

———, 1979b: Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1): pp. 339–359.

———, 1982: Logic for Equivocators. *Noûs*, 16(3): pp. 431–441.

———, 2009: Ramseyan Humility. In *Conceptual Analysis and Philosophical Naturalism*, D. Braddon-Michell and R. Nola, eds., pp. 203–222. Cambridge, Mass.: MIT Press.

Lipton, P., 1990: Contrastive explanation. *Royal Institute of Philosophy Supplement*, 27(1): pp. 247–266.

List, C., 2008: Which Worlds are Possible? A Judgment Aggregation Problem. *Journal of Philosophical Logic*, 37(1): pp. 57–65.

Millikan, R.G., 1984: *Language, Thought and Other Biological Objects*. Cambridge, Mass.: MIT Press.

Pincock, C., 2007: A Role for Mathematics in the Physical Sciences. *Noûs*, 41(2): pp. 253–275.

Powers, L.H., 1978: Knowledge by Deduction. *The Philosophical Review*, 87(3).

Price, H., 1983: Does 'Probably' Modify Sense? *Australasian Journal of Philosophy*, 61(4): pp. 396 - 408.

Putnam, H., 1975: Philosophy and our Mental Life. In *Mind, Language, and Reality: Philosophical Papers, Vol. 2*, pp. 291–304. Cambridge: Cambridge University Press.

Raiffa, H. and Schlaifer, R., 1961: *Applied statistical decision theory.* Boston: Harvard University Press.

Ramsey, F., 1931: General propositions and causality. In *The Foundations of Mathematics and other Logical Essays*, R.B. Braithwhaite, ed. London: Kegan Paul.

Rayo, A., 2011: A Puzzle About Ineffable Propositions. *Australasian Journal of Philosophy*, 89(2): pp. 289–295.

van Rooy, R., 2004: Utility, informativity and protocols. *Journal of Philosophical Logic*, 33(4): pp. 389–419.

Schroeder, M.A., 2008a: *Being For: Evaluating the Semantic Program of Expressivism.* Oxford University Press.

———, 2008b: How expressivists can and should solve their problem with negation. *Noûs*, 42(4): pp. 573–599.

Skyrms, B., 1977: Resiliency, Propensities, and Causal Necessity. *The Journal of Philosophy*, 74(11): pp. 704–713.

———, 1980: *Causal Necessity.* New Haven: Yale University Press.

———, 1990: *The Dynamics of Rational Deliberation.* Cambridge, Mass.: Harvard University Press.

Soames, S., 1987: Direct Reference, Propositional Attitutes, and Semantic Content. *Philosophical Topics*, 15(1): pp. 47–87.

Sober, E., 1998: Black Box Inference: When Should Intervening Variables Be Postulated? *The British Journal for the Philosophy of Science*, 49(3): pp. 469–498.

Stalnaker, R., 1968: A theory of conditionals. *Studies in logical theory*, 2: pp. 98–112.

———, 1973: Presuppositions. *Journal of Philosophical Logic*, 2(4): pp. 447–457.

———, 1984: *Inquiry*. Cambridge, Mass.: Bradford books, MIT Press.

———, 1991: The problem of logical omniscience, I. *Synthese*, 89(3): pp. 425–440.

———, 2002: Epistemic Consequentialism. *Aristotelian Society Supplementary Volume*, 76(1): pp. 153–168.

Steiner, M., 1998: *The applicability of mathematics as a philosophical problem*. Cambridge, Mass.: Harvard University Press.

———, 2005: Mathematics — Application and Applicability. In *The Oxford Handbook of Philosophy of Mathematics and Logic*, S. Shapiro, ed., pp. 625–650. Oxford: Oxford University Press.

Strevens, M., 2004: The Causal and Unification Approaches to Explanation Unified—Causally. *Noûs*, 38(1): pp. 154–176.

Sturgeon, N., 1988: Moral Explanations. In *Essays on Moral Realism*, G. Sayre-McCord, ed. Cornell University Press.

Sturgeon, S., 2008: Reason and the Grain of Belief. *Noûs*, 42(1): pp. 139–165.

Swanson, E., 2006: *Interactions with Context*. Ph.D. thesis, Massachusetts Institute of Technology.

Unwin, N., 1999: Quasi-Realism, Negation and the Frege-Geach Problem. *The Philosophical Quarterly*, 49(196): pp. 337–352.

———, 2001: Norms and Negation: A Problem for Gibbard's Logic. *The Philosophical Quarterly*, 51(202): pp. 60–75.

Wallace, R.J., 2001: Normativity, Commitment, and Instrumental Reason. *Philosophers' Imprint*, 1(4): pp. 1–26.

White, R., 2005: Explanation as a Guide to Induction. *Philosopher's Imprint*, 5(2).

Williams, J.R.G., forthcoming: Counterfactual triviality: A Lewis-impossibility proof for counterfactuals. *Philosophy and Phenomeno-*

logical Research.

Woodward, J., 2010: Scientific Explanation. In *The Stanford Encyclopedia of Philosophy*, E.N. Zalta, ed. Spring 2010 edn.

Yablo, S., 1992: Mental Causation. *The Philosophical Review*, 101(2): pp. 245–280.

———, 2001: Go Figure: A Path through Fictionalism. *Midwest Studies in Philosophy*, 25(1): pp. 72–102.

Yalcin, S., 2008: *Modality and Inquiry*. Ph.D. thesis, Massachusetts Institute of Technology.

———, forthcoming: Nonfactualism about Epistemic Modality. In *Epistemic Modality*, A. Egan and B. Weatherson, eds. Oxford University Press.